This is to certify that the

dissertation entitled

The Impact of Test Consequences and Response Format on Performance

presented by

Christine E. DeMars

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Measurement and Quantitative Methods

William C. Mehrens
Major professor

Date 1-6-98

THE IMPACT OF TEST CONSEQUENCES AND RESPONSE FORMAT ON PERFORMANCE

By

Christine DeMars

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods

1998

ABSTRACT

THE IMPACT OF TEST CONSEQUENCES AND RESPONSE FORMAT ON PERFORMANCE

By

Christine DeMars

Students generally perform higher on tests that have some consequences. This study examined whether the performance difference between high- and low-stakes tests remained constant across response formats, genders, and ethnic groups. Also, this research explored whether any of these factors were associated with non-response, and compared the fit of item estimates based on student responses under low-stakes to the response patterns under high-stakes. Data were obtained from pilot and operational administrations of the science and mathematics sections of the Michigan High School Proficiency Test (HSPT).

Results showed that students were more likely to respond to constructed response items, and to score higher on the overall test, when the test had consequences. Whites were more likely to respond than Blacks, and Whites scored higher on the test. Girls were more likely to respond than boys, but boys scored higher overall on the science test and there were no gender differences in overall scores on the math test. Increased test stakes increased performance on constructed response items more than on multiple choice items,

and the gender difference in performance changed with the response format (boys scored higher on the multiple choice section, while girls scored higher on the constructed response section). The ethnic by format interaction depended on the subject matter and test form, but tended to be small. In math, pilot item estimates fit the operational data better when omitted items were treated as not-administered, rather than incorrect, during the item estimation. They also fit girls better than boys, and multiple choice items better than constructed response items. In science, where the fit statistics were better and response rates were higher, differences in fit were small.

ACKNOWLEDGMENTS

Writing these acknowledgments is among the last of my tasks in completing this dissertation, but those who have helped me along the way certainly are not last in my thoughts.

I would like to thank my committee members for all their assistance. Dr. Betsy Becker gave me especially detailed feedback on my proposal, and from her course I learned the skills needed for the IRT portion of this project. Dr. Susan Phillips kindly stepped-in at the last moment to join the committee as Dr. Becker prepared to leave for a sabbatical. Dr. Steve Raudenbush gave me guidance in defining my models and interpreting my results. Dr. Neal Schmitt provided a fresh perspective from industrial/organizational psychology. Dr. Chris Schram, as an alumnus of my program and current staff member of MEAP, was a helpful link to the "real-world" context of my study. All my committee members provided useful feedback, and I appreciated their positive, helpful attitudes at committee meetings.

Most of all, I am grateful for the assistance of my advisor and dissertation director, Dr. William Mehrens. He guided me throughout the dissertation process, as well as earlier "milestones" in my doctoral program. He always found time to discuss my progress and encourage me. I knew I could count on him for prompt, helpful (and tactful) feedback on my work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Purpose

The purpose of this study was to examine how responses to items on the science and math sections of the Michigan High School Proficiency Test (HSPT) changed as the stakes of the test changed, and how (or if) these changes were associated with the response format of the items and the examinees' gender and ethnicity. The test was initially administered in the 1994-95 academic year under low-stakes (essentially no-stakes) conditions, as a pilot test (final field test). In the spring of 1996 and 1997, the test was taken by the next cohorts of students under high-stakes conditions--students who did not score satisfactorily would not be eligible for a state-endorsed diploma. Both constructed response and multiple choice items were on the test. Two types of changes in response were studied: changes in the quantity of responses (response rate) and changes in the quality/correctness of responses (item scores). Also, the fit of the operational data to the parameter estimates of the pilot data was examined, as a measure of the accuracy of the pilot estimates.

Need

Test developers and researchers want estimated item difficulties from pilot tests to reflect the relative difficulties the items will have under operational conditions. If one type of item changes more than another, the test developers will not get the mix of item

difficulties they intended. If the performance of one demographic group changes more than another, estimates of group differences will not be accurate, which makes it difficult to assess the impact of cut scores or the educational needs of the groups. Further, if there is an item by group by testing condition interaction, estimates of differential item functioning will not be the same.

Also, pilot tests share similarities with other low-stakes tests, such as the National Assessment of Educational Progress (NAEP) and various exams conducted by the International Education Association (IEA). On these examinations, students are informed that their scores will be anonymous and they receive no individual feedback. These tests are often even more visible than the tests which have stakes on an individual level. Findings about pilot tests could generalize to these other important low-stakes situations. Low-stakes tests may underestimate student performance, and they may not result in accurate estimates of group differences or response format differences. Accurate estimates from these tests are important because results from these tests may be used in making policy decisions.

The Study

The influences of response format, gender, and ethnicity on changes in performance from low to high stakes testing conditions were the focus of the present study. The math and science sections of the Michigan High School Proficiency Test (HSPT) were investigated here. This test was administered to a sample of 11th graders as a pilot test (low stakes) during the 1994-95 school year. This exam was later administered to all 11th graders in the state during the following years, leading to state high school diploma endorsements (high stakes). The test responses of students in the schools which

participated in the piloting of the 1996 science form and 1997 math form were studied here (these were the first years, respectively, that most students needed this test for the diploma endorsement--in 1996, many of the tested students had earned math endorsements from a previous test). The science test form had 8 written constructed response items and 34 multiple choice items (there were 8 additional multiple choice items which were different on the operational test--these served as linking items across forms). The math form had 6 constructed response items and 32 multiple choice items (again, there were 8 additional items which were different on the pilot and operational forms). Items in both formats were intended to assess some higher level cognitive skills as well as basic content knowledge. Under both low and high stakes conditions, students were allowed as much time as they needed to complete the test (according to the administration instructions). Blacks[1] were the only ethnic minority members who participated in large numbers in the pilot study, so this was the only ethnic minority group studied here.

Existing literature suggests motivation and performance should be higher on high stakes exams. Wolf, Smith, and Birnbaum (1995) found students who could be placed in remedial classes based on test scores had greater motivation and omitted fewer items than students facing no consequences. Wolf and Smith (1995) and Wolf, Smith, and DiPaolo (1996) recorded considerably higher motivation scores (a difference of about 1.5 standard deviations) in consequential than non-consequential test conditions. Arvey, Strickland, Drauden, and Martin (1990) measured higher motivation in job applicants taking a screening test than in current employees.

---

[1] The label "Black" was used here because that was the option listed on the student identification sheet. Some students who marked this option may prefer the term "African American".

While motivation increases on high stakes tests, anxiety may increase as well. This could decrease test performance, because highly anxious students tend to score lower on tests (Crooks, 1988; Hembree, 1988; Wolf & Smith, 1995; Wolf, Smith, & DiPaolo, 1996). On average, though, higher consequences lead to higher test scores (Burke, 1991; Jennings, 1953; Rothe, 1947; Taylor & White, 1981; Wolf & Smith, 1995; Wolf, Smith, & DiPaolo, 1996). For the majority of students, then, it appears that the motivating effects of consequences are stronger than the anxiety-provoking effects.

This study differed from previous research in that it focused on how characteristics of the items and of the test-takers impacted the degree of the effect of motivating consequences. This study assessed whether test consequences had a greater (or lessor) effect on performance on constructed response items than on performance on multiple choice items. This study also explored whether differences in test consequences affected males and females to the same extent, and whether the degree of change was similar for Blacks and Whites.

Given Freund and Rock's (1992) findings that males and Blacks appeared to be less motivated on a low-stakes test, and Karmos and Karmos' (1984) findings that the correlation between attitude and test scores was higher for males, a tentative hypothesis was that the performance of males and Blacks would increase more as the stakes increased. This would also be supported by Kiplinger and Linn's (1992) evidence that the scores of Blacks increased slightly more than the scores of Whites as consequences increased, though the difference was small. These findings also suggested that scoring omitted items as zero rather than treating them as "not administered" would have a larger impact on ability estimates for Blacks and males.

Regarding response format, nonresponse has been high for low-stakes constructed response items (Badger, 1989; Freund & Rock, 1992). Motivation has been found to be lower on constructed response items compared to multiple choice items under low stakes (Sundre, 1996; Wainer, 1993). If this "motivation gap" narrowed with the external motivation of high stakes, performance differences would likely decrease as well. Findings that performance on "mentally taxing" items increased more than performance on less taxing items under higher stakes (Wolf, Smith, & Birnbaum, 1995) also suggest that response format effects (on both response rate and test scores) would decrease under higher stakes. Findings on ethnic-group differences on constructed response items compared to multiple choice items are mixed (Badger, 1995; Bond, 1995; Feinberg, 1990; Klein, Jovanovic, Stecher, McCaffrey, Shavelson, Haertel, Solano-Flores, and Comfort, 1997; Linn, Baker, & Dunbar, 1991). If some of the ethnic-group differences were due to one group having especially low interest or motivation on low stakes tests, the differences would decrease under higher stakes, but differences due to curricular/instructional differences would not be affected by the stakes. Specifically, the following questions were addressed:

(1) What effects do test consequences, response format, gender, and ethnicity have on response rate?

(2) What effects do test consequences, response format, gender, and ethnicity have on test scores?

(3) How well do item parameter estimates based on pilot data fit the operational data, and how do pilot estimates based on treating omitted items as not-administered compare to estimates based on treating omitted items as incorrect?

To analyze non-response, students' responses (coded 0 for non-response and 1 for response) on the constructed response items (response rates were universally high on the multiple choice items, so they were not included) were analyzed in a hierarchical generalized linear model. The responses were assumed to be binomially distributed; a transformation of the probability of responding (the log-odds of response) was used as the outcome variable, because the error variance of the dichotomous outcome would be neither normally distributed nor independent of the underlying probability of response. At the lowest level, within students, the probability of response was modeled as a function of the students' general tendency to respond. At the between-student level, several factors potentially influenced the student's average response tendency: the test stakes for that student, the student's gender, and the student's ethnic group. At a higher level, school factors impacted the student factors as well as the average response tendency within the school. One of these factors was the proportion of students in the school who identified themselves as non-White (which, among other things, serves as a proxy variable for SES). Other school factors were not modeled here, but the variance of random school effects was assessed, so that the variance at the other levels could be more accurately estimated.

For the analysis of the item scores, neither the log-odds model nor the usual linear model was appropriate, because some of the items were dichotomous while others were on a multi-point scale. Instead, a one-parameter item response model (partial-credit model for the constructed response items) was used to estimate two composite ability scores for each student, one based on the constructed response items and one based on the multiple choice items. The one parameter model was used because it is the model used to equate and to score the HSPT. The item parameters were estimated from the operational

(high stakes) administration, because effort was more likely to be stable across items in this group (thus effort was less likely to be a second dimension underlying performance than it might be if both samples were combined to estimate item parameters). These parameter estimates were used to estimate the two abilities for each student in both samples. The parameters for all items were calibrated simultaneously for both sets of items to place them on the same metric, so for the calibration students the two ability estimates were the same, on average. However, if the relative difficulty of the items in the two scores was not constant across subgroups of students, the ability estimate for one format would be consistently higher in the affected subgroup.

These scores were the dependent variables in a hierarchical linear model. At the first level, response formats (two) were nested within students. The error variance was based on the standard errors of measurement estimated from the IRT analysis, which were different for each scale for each student (the error variance was not homogeneous). At the next level, the student factors were those used in the non-response analysis: the test stakes for that student, the student's gender, and the student's ethnic group. A unique term for uncontrolled variance between students was also in the model. At the school level, school-effects again were modeled to depend on the proportion of students who were minority members. The variance of the school means (conditional on the means predicted from the student ethnic composition) was estimated to allow for accurate estimation of error variance at the other levels.

At the lowest-level, the average format coefficient showed the average performance difference between the two formats. At the student level, the average coefficients estimated the average performance difference within schools due to the stakes

of the test, ethnic group, and gender as well as the impact of each of these factors on format differences. At the school level, the association between school percent-minority and each of the factors at the lower levels (including the school intercept) was estimated.

The analysis described above used item estimates based on the high-stakes data. In a further analysis, item parameters were re-estimated from the low-stakes data, to see how well these estimates will fit the response patterns of the high-stakes students. Estimates were made twice under low-stakes conditions, once treating omitted items as incorrect and again treating omitted items as not-administered. Parameters were estimated separately for males and females.

These item estimates from the pilot data were then used to re-estimate the abilities of the high-stakes students. An estimate of ability was obtained for each student, using gender-specific item estimates, once using the item estimates based on scoring omits as zero, and again using the item estimates based on treating omits as not-administered. The ability estimates were based on all items, with omitted items scored as wrong (regardless of how omitted items were treated in the item estimation). Based on this ability estimate, the fit of each student's responses to the item parameters was calculated, using the Drasgow, Levine, and Williams (1985) standardized appropriateness index for polychotomous data. This index was used to flag students who had particularly poor fit. The number of misfitting students was compared across item-estimation conditions and genders. If there were more boys than girls who had poor fit, for example, it would mean that the pilot item estimates were less accurate for boys. The information about which way of treating omits produced the most accurate estimates can be used by those who make decisions of how to treat omits for item estimation for low-stakes tests.

The appropriateness index measures "person-fit". "Item-fit" was also of interest, because it is the item parameters which are used in equating and in some types of Differential Item Functioning (DIF) estimates. Generally, if person-fit is poor then item-fit will also be poor, but it is possible for a few items to misfit without highly affecting the overall person-fit. To measure item-fit, the OUTFIT statistic (Wright & Masters, 1982) was calculated. For this measure, the standardized difference between each person's observed and expected score (based on ability and item difficulty) was found for each item, and these standardized differences were summed across people. Item difficulty estimates, again, were based on the low-stakes responses while the responses used in the test of fit were the responses of the high-stakes students.

The average OUTFIT was calculated for each gender group, estimation condition (method of treating omits), and response format (constructed response or multiple choice). These means were compared to see if gender, way of treating omits, or response format affected the fit. Especially large values for individual items (1.5 or more--50% greater than average) were also noted.

If there were a stakes by format performance interaction, the measures of fit would tend to be worse for the items which showed the greater performance difference between the two test conditions. If one gender group showed greater performance differences when the stakes of the test changed, the fit could be worse for this group. If there were a gender by stakes interaction in tendency to omit items, the fit would be particularly poor when omitted items were treated as incorrect for the group which increased its response rate more.

CHAPTER 2

REVIEW OF THE LITERATURE

How do the consequences of a test affect student performance, and do the consequences affect students of different genders or ethnic groups differently? Do the consequences have a greater impact on performance on constructed response or multiple choice items? Previous research has touched on some of these issues.

Effects of Consequences on Test Scores

Examinees generally score higher under high stakes conditions. An early example of this phenomena was described by Rothe (1947). On each of four tests (two each for laundry workers and machine-shop workers), the scores of job applicants (high stakes) were higher than scores of current workers (low stakes). The differences ranged from about 0.20 standard deviations to greater than one standard deviation. Rothe (1947) ruled out several alternative explanations: the applicants were younger (so they would be expected to score higher on certain tests, especially speeded tests), but two of the four tests were unrelated to age; the laundry plant applicants were female so they were not test-wise from the military (as male job applicants were likely to be); the new applicants were applying for the same types of jobs as the current workers (there was not an imbalance of clerical v. shop workers); and there was not a recent drop in applications, which could indicate self-selection. Rothe concluded the applicants did better because they were more motivated. Similarly, Jennings (1953) found that supervisors who thought their test scores

would be used for deciding promotions scored more than one standard deviation higher than those who were told the testing was for research purposes.

Other examples are more recent and more relevant to educational testing. Wolf and Smith (1995) administered two counterbalanced forms of a class exam to undergraduates who were aware which test would be graded. On average, students scored 0.26 standard deviations better (64% compared to 61%) on the consequential exam, though some individuals (about 1/3) scored better on the nonconsequential form. Similarly, Wolf, Smith, and DiPaulo (1996) found undergraduates scored an average of 0.44 standard deviations better under consequential conditions. Taylor and White (1981) increased the consequences for Title I second-graders taking the Stanford Achievement Test by offering the students money for improving their scores. This reinforcement led to an increase of more than one standard deviation in reading scores.

Similar effects have been found when stakes have increased in non-experimental conditions. Burke (1991) demonstrated that scores on the NAEP in Louisiana and Maryland improved when these states began using the test for high school graduation (Burke also suggested increased curricular alignment due to this use of the test could play a large role). However, when NAEP mathematics items were added to Georgia's state tests, scores did not improve on one of two forms and increased by 0.18 standard deviations on the other, a difference similar in magnitude to differences found on other NAEP tests due to context changes (Kiplinger & Linn, 1992). Though these tests had some consequences at the school level, they had no consequences at the individual student level so smaller changes would be expected (in fact, a focus group of 12th graders recommended increasing motivation by not informing students the test had no individual

consequences). Even in studies where the group mean increased as the consequences increase, not all individual student scores increased. In Jennings' (1953) sample, the rank order of supervisors tested under consequence and no-consequence conditions was unstable; some of those with high scores under one condition had low scores under the other condition. In Wolf and Smith's (1995) sample, about 1/3 of students scored higher under no consequences.

Effects of Consequences on Motivation

Why is performance higher on high-stakes tests? One explanation is that test consequences increase motivation, which in turn increases performance. In the present study, the measure of motivation is response rate, a somewhat muddy indicator because non-response may be due to lack of knowledge rather than lack of motivation. Previous findings about the effects of consequences on motivation are described in this section, and the link between motivation and performance is the topic of the next section.

Motivation seems to be low on many low-stakes exams. About 36% of a group of 8th graders who participated in the NAEP thought it was "not very important" or only "somewhat important" to do well on the test (Kiplinger & Linn, 1992). Further, 28% reported they had tried "not at all hard" or only "somewhat hard" on the test. In preliminary studies, external consequences increased the scores of eighth graders, and the information that schools or students would be compared had a greater effect than small amounts of money. Twelfth-graders, in a focus group, also reported very low levels of motivation (Kiplinger & Linn, 1992). Given these low levels of motivation, it might be expected that some students would just fill-in bubbles on the answer sheet to complete the test quickly and without effort. Freund and Rock (1992) developed an algorithm to

identify students who were suspected of using some type of pattern to mark their answer sheet (patterns would tend to lead to similar differences between the locations of any two responses). Eighth and 12th graders who scored more than 1.5 standard deviations below the eighth-grade mean for this measure were identified as "pattern markers" on the NAEP. More males than females, and a greater proportion of Hispanics and Blacks than Whites, met this criterion. Freund and Rock (1992) suggested it might be better to delete the responses of pattern-markers before item estimation, because these responses would add "unnecessary" variance.

Motivation has been found to be low on other tests as well. A group of 45 college students whose scores decreased between their freshmen and sophomore years on a low-stakes test (intended to assess university goals) participated in interviews concerning the exam (Olsen & Wilson, 1991). The students said they were not very serious about the test and did not approach it like a "real" classroom test. When asked how much various things would motivate them, the students rated personal stakes (free electives for high scores, additional courses if low scores, and even personal feedback on performance) as more motivating than consequences for the university (government funding). This might suggest that high school and middle school students (such as those in Kiplinger & Linn, 1992) would be unlikely to be motivated by school or district level stakes.

On the Michigan Educational Assessment Program (MEAP) reading tests, which previously had few individual stakes, motivation decreased with grade level (Paris, Lawton, & Turner, 1992; Paris, Turner, Lawton, & Roth, 1991). Students were asked if they had read all the reading passages on the test and if they had checked their answers. High school students were less likely than younger students to have done these things.

Greater proportions of high school students said they had just filled-in some of the answers without trying to choose the correct answer. The survey also asked students how much they cared about their scores, and how much they thought their parents and teachers cared about their scores; high school students responded more negatively than younger students. These findings were similar to the results from two broader surveys of students in four states: Older students had more negative attitudes and engaged in poorer test-taking behaviors.

These results show motivation is low on low-stakes tests, but they do not necessarily imply that motivation is higher on high-stakes tests. Other studies, though, have compared motivation under high and low (no) consequence conditions and have found higher motivation when the test has individual consequences. Wolf, Smith, and Birnbaum (1995) studied a test given in New Jersey to 10th graders to assess need for individual remediation and to 11th graders for high school graduation. At the first administration, though, few 11th graders were taking the test for graduation because they had passed a previous graduation exam as 9th graders. The 10th graders rated their effort on the test significantly higher than the 11th graders did, and 10th graders omitted fewer items (though omit rates were low for both groups). Similarly, Wolf and Smith (1995) and Wolf, Smith, and DiPaulo (1996) collected motivation scores for college students who completed two exams, only one of which contributed to their course grades. In both studies, the motivation scores were about 1.5 (1.45 and 1.58) standard deviations greater for the test which was included in the course grade. Parallel results have been found on personnel tests. Arvey, Strickland, Drauden, and Martin (1990) measured test motivation for a job screening test for highway maintenance workers. Job applicants had higher

motivation scores (greater than one standard deviation) than current workers. In another sample of applicants for a county financial position, they found Whites had higher motivation scores than Blacks on the application test, and controlling for test attitudes lowered the association between ethnic group and test performance.

Effects of Motivation on Performance

Motivation is relevant to testing issues because increased motivation tends to lead to improved test performance. There have been at least two meta-analyses summarizing this relationship (Multon, Brown, & Lent, 1991; Uguroglu & Walberg, 1979). Combining findings from 232 samples, Uguroglo and Walberg (1979) found an average correlation between motivation (including measures of self-concept and locus of control) and achievement (both test scores and course grades) of .34. This association tended to be higher for older students and for class grades compared to achievement tests (and lowest for general ability tests), but it did not vary significantly by gender. Multon, Brown, and Lent (1991) also found the correlation between feelings of self-efficacy (their model proposed that self-efficacy increased motivation) and test performance increased with grade level and was higher on classroom and basic skills tests than on standardized achievement tests. Their average unbiased correlation, for 38 samples, was .38. Multon, Brown, and Lent also synthesized 18 samples from studies of the relationship between self-efficacy and the number of items answered on a test, finding an average correlation of .48. In many of these studies, though, it is difficult to know whether the more motivated students did better because of high motivation/confidence, or if the higher motivation/confidence was due to the students' knowledge of their own high abilities. Also, Multon, Brown, and Lent did not show how highly related self-efficacy and

motivation were in the samples they synthesized; the link between self-efficacy and motivation was based on theory and results from other studies.

The relationship between performance and motivation may vary by gender. Though the meta-analysis of Uguroglo and Walberg (1979) showed no average gender effects on the correlation, Karmos and Karmos (1984) found the correlation was higher for males than for females. Females, however, had more positive attitudes toward tests, and this gender difference was more apparent at higher grades (students from grades 6-9 participated in the study). The attitudes of males had a greater variance. However, Brown and Walberg (1993) found no interaction between gender and motivation when motivation was operationalized as an experimentally-manipulated variable rather than a characteristic of students, and their sample included middle school students (as Karmos & Karmos did) as well as elementary students. For both males and females, motivating instructions (informing students their school would be compared to other schools and their scores would be used in evaluating their teachers) increased achievement test scores an average of 0.30 standard deviations.

If external motivation is high enough, test-takers may report only small differences in motivation, depressing the correlation between test scores and motivation scores. Arvey, Strickland, Drauden, & Martin (1990) measured the test attitudes (including motivation) of job applicants and current employees (highway workers) taking several job-related tests. The variance in motivation scores was over five times as high in the employee group compared to the applicant group. The correlations between motivation scores and test scores were small but significant and consistent in the employee group (r = .24, .25., and .23 on three tests). In the applicant group, where motivation scores were

uniformly high (though test scores were more variant), the correlations were not significantly different from zero (-.01, .04, .10). In a sample of applicants in another occupation (county financial workers), correlations of test scores with motivation scores were small; the largest correlation, between a simulated work sample and motivation, was .20. There were no comparisons with current employees with this classification, nor was the variance on the motivation scale reported.

Some tasks may be more motivating because test-takers find them more interesting. For example, computer adaptive tests are more interesting than paper-and-pencil versions for many examinees. The computer adaptive version of the Armed Services Vocational Aptitude Battery (ASVAB) was described as less boring, more challenging, and more interesting by a sample who took both the computer and standard versions (Arvey, Strickland, Drauden, & Martin, 1990). These test-takers also reported working harder on the computer version (in effect, they were more motivated). A large sample of recruits took both computer-adaptive and paper-pencil versions of the ASVAB, as well as the Test Attitude Survey (TAS) after each version; these recruits had higher average attitude scores associated with the computer version (Arvey, Strickland, Drauden, & Martin, 1990). Another study of randomly equivalent groups of recruits taking the test under no-stakes conditions showed that the equating of the computer and paper versions of the tests did not generalize to operational conditions (Segall, 1997). Under no-stakes conditions, the computer adaptive test appeared easier, relative to the paper test, than it did under operational conditions. Essentially, scores on the paper version increased more as the stakes increased, changing the relationship between the formats.

Children tend to score higher on reading comprehension when the topics of the reading passage interest them (Asher, 1979; Asher, 1980; Bernstein, 1955; Stevens, 1979). These findings hold whether the students' interest is measured on the specific passages after reading (Bernstein, 1955) or on the general topics of the passages in a seemingly unrelated session at some point before the test (Asher, 1979; Asher, 1980; Bernstein, 1955; Stevens, 1979). Findings regarding the interaction of gender and interest on performance are mixed. Asher (1980) and Bernstein (1955) found that testing students on topics they were interested in increased the scores of boys more than girls. However, there was no interaction between gender and interest in the studies of Asher (1979) and Stevens (1979). Only one of these studies looked for a possible interaction between ethnicity and interest; Asher (1979) showed that the scores of Blacks increased to the same degree as the scores of Whites when students' comprehension scores were compared on topics they rated as high and low on interest.

An interesting finding in this area is that the effects of interest are less when there are some consequences for students. Asher (1980) found that scores on high interest topics were not significantly higher than scores on low interest topics when students were offered external incentives. This suggests that interest is more motivating in the absence of other motivation and may not play a large role on high stakes tests where students are already motivated to earn high scores. Scores on low-interest tasks or items, then, may increase more from pilot to operational test than scores on high-interest tasks. This is relevant to the present study if interest is associated with response format on the HSPT.

Response Format

Constructed response items are often viewed as performance tests, or as being closer to authentic performance tests than multiple choice items. In writing, for example, writing an essay is a sample of the desired performance (writing) to be measured. In mathematics, the target performance might be problem-solving in ill-structured situations, which could be operationalized as a written task where students explained their reasoning in interpreting and solving the problem. In science, the desired performance might be designing and conducting appropriate experiments; written responses in which students explained how they would do this or interpreted results supplied in the test could be considered a less expensive simulation. More authentic performance tests in science have also been proposed for large-scale use: Gao, Shavelson, & Baxter (1994) and Shavelson, Baxter, & Pine (1992) reported research on field tests of science assessments where students designed and carried out scientific investigations.

Effects of Response Format on Group Differences

Some proponents of constructed response tests suggest that ethnic and/or gender differences will be lower on constructed response tests. Messick (1994) explained that proponents of this position contend that there is "less construct under-representation and construct-irrelevant method variance" in performance tests. If there were no group differences on the construct, then there should be fewer differences on such tests (if the claims for their validity were assumed). On a test for skilled metal-workers, for example, there were no ethnic group differences in the scores on the actual products, but there were fairly large differences (over one standard deviation on the total score) on a paper-and-pencil test covering the same machines (Schmidt, Greenthal, Berner, Hunter, & Seaton,

1977). On a test of situational judgment in the workplace for blue-collar workers, Black/White performance differences were smaller when the scenarios (prompts) were displayed on videotape rather than described in writing (Chan & Schmitt, 1997).

However, academic performance tests, especially large-scale tests, often involve a *written* product (even if there is some manipulation of materials), which might add irrelevant variance if written communication were not part of the target construct. In the Chan and Schmitt study (1997), controlling for the reading skills by testing method interaction (level of reading skills had a larger impact on test scores for the written prompt method) decreased the race by test method interaction. In most large-scale tests in education, though, reading/writing skills are likely to be at least as influential in "performance" tests as in multiple choice tests. Also, there might be larger group differences (due perhaps to educational differences) on the types of constructs (such as science or math knowledge) measured in education--removing "construct irrelevant" variance would not change this.

One series of tests with school-level stakes, the Massachusetts Educational Assessment Program, which included both multiple choice and written constructed response items, *did* find smaller ethnic group differences on constructed response items (Badger, 1995). In eighth grade mathematics, the scaled scores of Hispanic and Black students were higher on the constructed response items than the multiple choice items, bringing them closer to the scores of White and Asian students, and patterns were reportedly similar in other grades and subject areas. Many educators have emphasized features which need attention if alternative assessments are to be equitable (Darling-

Hammond, 1995; Roeber, 1995; Winfield, 1995), but with no concrete evidence yet that the resulting tests are more equitable than multiple choice tests.

However, other studies have found ethnic group differences on constructed response tests are likely to be at least as large as group differences on multiple choice tests. On science items field tested for the California Learning Assessment System (CLAS), ethnic differences (White/Black, Anglo/Hispanic) were relatively similar on multiple-choice items and written constructed response items following hands-on activities, especially among fifth and sixth graders (Klein, Jovanovic, Stecher, McCaffrey, Shavelson, Haertel, Solano-Flores, and Comfort, 1997). In ninth grade, the White-Black difference was 0.51 standard deviations on the constructed response section and 0.85 on the multiple choice section; the White-Hispanic difference was more similar across formats (0.61 and 0.73). On the California Bar Exam, a written section added in 1984 did not decrease the differences between the average scores of whites and minorities, and the rank order of students was similar across the essay, performance, and multiple choice sections (Feinberg, 1990). Differences between Blacks and Whites on the 1988 NAEP writing tests (which were essay tests) were of about the same size, in standard deviation units, as differences on the reading tests (primarily, though not entirely, multiple choice) (Linn, Baker, & Dunbar, 1991). Also on the NAEP, Bond (1995) reported that differences between the scores of Blacks and Whites were greater on extended response items than on multiple choice items. On performance tests developed in Great Britain, gender and ethnic group differences increased on the constructed response items (Nuttall & Goldstein, 1990, cited in Shepard, 1993).

Gender differences on the Advanced Placement exams (high stakes) depend on response format. In most science areas (Biology, Chemistry, Physics B, Physics C-E&M, Physics C-Mech) the gender difference (favoring males) is smaller on the constructed response items (Bridgeman, 1989; Bridgeman & Lewis, 1994; Schmitt, Mazzeo, & Bleistein, 1991). This pattern of gender differences was found in all ethnic groups tested in large numbers in biology (White, Asian American, and Black) and chemistry (White and Asian American). Bolger and Kellaghan (1990) found the same pattern with Irish high school students. Klein, Jovanovic, Stecher, McCaffrey, Shavelson, Haertel, Solano-Flores, and Comfort (1997) found fifth and sixth grade girls scored significantly higher than boys on written constructed responses to hands-on science activities, while gender differences were nearly zero on a multiple choice science test (ITBS). By ninth grade, boys scored significantly higher on a multiple choice science test, while girls scored somewhat higher on the hands-on constructed-response test. On the pilot (low stakes) administration of the Michigan High School Proficiency Test, DeMars (in press) found a similar pattern in mathematics and science (gender differences favoring males were greater on the multiple choice section), but gender differences in most of the ability range were small. Also, gender differences were relatively stable across format in the calculus (Calculus AB and Calculus BC) and computer science AP Exams (Schmitt, Mazzeo, & Bleistein, 1991).

The NAEP is a low-stakes test, and Freund and Rock (1992) found higher rates of suspected "pattern-marking" by Blacks and Hispanics on the multiple choice section. This suggests that minorities have especially low motivation on low stakes tests, and if this affects constructed response items more than multiple choice items, the scores of

minorities may increase relatively more on constructed response items than multiple choice items under high stakes conditions where students are motivated on both types of items.

Because constructed response items take longer to complete than multiple choice items, fewer tasks are sampled. Linn, Baker, and Dunbar (1991) raised the possibility that this would "make it harder to achieve balance with regard to group differences in prior knowledge" (p. 18). When many items are utilized, group differences are more likely to cancel out (assuming the item-specific differences are incidental to the primary construct). Feinberg (1990) explained: "Compared to multiple-choice tests of similar length, written exams more arbitrarily emphasize one topic or another with which a student may (or may not) be familiar" (p. 30). Similarly, Messick (1994) explained that "contextualizing" items (a purported advantage of using fewer, longer items, including constructed response items) will affect individual students differently, depending on their familiarity with the context. He noted:

> We should not take it for granted that a richly contextualized assessment task is uniformly good for all students . . . contextual features that engage and motivate one student and facilitate his or her effective task performance may alienate and confuse another student and bias or distort task performance. (p. 19)

While constructed response items do not necessarily involve greater contextualization, given the greater time devoted to an individual constructed response item compared to a multiple choice item, there is greater opportunity for contextualization and more "authentic" tasks.

Shepard (1993) also pointed out that assessments more closely aligned with recommended curriculum may result in greater ethnic differences than more general assessments, if minority groups have less access to the assessed curriculum. Johnson

(1995) voiced similar concerns. Messick (1994) recommended studying the possibility of such curricular and instructional differences. Variance in time allocated to each content area can be large, even within a sample of low-SES schools (Winfield, 1995), so it is difficult to generalize about opportunity to learn for groups of students. Minority students are over-represented in the lower tracks of school, where some say they are more likely to "experience instruction geared only to multiple-choice tests" (Darling-Hammond, 1995, p. 96) presumably, from the additional context, meaning multiple choice tests which measure lower-level skills. Dreeban and Gamoran (1986) showed that controlling for time spent on reading instruction and individual aptitude reduced (essentially to zero) the effects of race and SES on first grade reading achievement, concluding that students in primarily Black schools (most of the schools studied were racially homogeneous) received inferior instruction. In Massachusetts, teachers in low-advantaged schools were less likely to think their students were prepared to answer items involving judgments, inferences, scientific procedures or mathematical reasoning (which were emphasized in constructed response items on the state tests), though their confidence in their students' preparation for factual or computational tasks was equivalent to the teachers' confidence in high advantaged schools (Badger, 1995). Similarly, students in low-advantaged schools were more likely to agree that "learning is mainly memorizing". The low-advantaged schools spent less money on science materials, and the math teachers in these schools were less likely to feel they had adequate equipment (calculators, computers, manipulatives). These results suggest curricular differences favoring the advantaged students. Therefore, it is surprising that minority and disadvantaged students did *relatively* better on the constructed response items, which purportedly measured more higher-level skills than the

multiple choice items (the scores of minority students were lower on both types of items, but the gap was somewhat smaller on the constructed response items).

Effects of Response Format on Motivation and Performance

One manifestation of low motivation is nonresponse. Conceptually, nonresponse seems to be more likely on constructed response items than on multiple choice items. It takes very little effort to mark an answer on a bubble grid; it is not even necessary to read the item. However, writing an answer (even one unrelated to the question) takes some thought and time. Nonresponse, though, could also signify the student knows nothing about the question. If a student has a very low level of knowledge, the student could guess randomly on the multiple choice items but not the constructed response items. The response pattern would be similar to that of an unmotivated student: random responses on multiple choice items and nonresponse on constructed response items. This pattern, then, while suggesting low motivation, is not a sure indication of lack of motivation.

Nonresponse or irrelevant response is common on low-stakes constructed response tests. The Massachusetts Educational Assessment Program, for example, has no student-level stakes. In one sample of students taking the mathematics portion, where students were asked to generate solutions to relatively ill-structured problems, many of the 8th and 12th graders simply left the items blank (Badger, 1989). The highest nonresponse rates were for an item which asked students to explain how adding a constant to each number in a set would affect the average (students were given concrete numbers to work with), and for an item which required students to explain how they would estimate the product of two numbers (a specific pair of numbers was supplied for students). Over 20% of the students left each of these items completely blank, not even supplying a numerical

answer (numerical answers with no explanations or irrelevant explanations were also common). On other constructed response items, non-response rates were about 8-15%.

On science constructed response items (completed by another sample of students at the same time), nonresponse rates (including irrelevant responses) for 8th and 12th graders ranged from 5% to 30% (Badger & Thomas, 1989b). A physics content item ("explain how a fuse works in a circuit") was most frequently left blank (30%). This might suggest students were omitting items because they lacked the specific content knowledge needed, but two process-type items also had nonresponse rates of nearly 20%.

In social studies (Thomas, 1989), response rates varied widely, with from 4% to 73% of 8th and 12th graders omitting items or supplying irrelevant answers. In reading (Badger & Thomas, 1989a), percentages of students giving no answer were reported separately from those giving irrelevant answers. On three 8th grade items, 6-10% of the students gave no answer, and another 6%-25% gave irrelevant answers. On one 12th grade item, 6% left the item blank and 12% gave irrelevant answers; on another item, 3% left it blank and 7% supplied an irrelevant answer. As reading generally demands less specific content knowledge than other subjects, such responses are more logically related to low motivation.

Nonresponse rates are also higher on constructed response items than on multiple choice items on the NAEP (Freund & Rock, 1992). Wainer (1993) citing Bock (1991) discussed how students taking the California Assessment appeared to be more motivated on the multiple choice items than on the constructed response items.

Response rates were also fairly low in a sample of students who participated in the second follow-up of the National Education Longitudinal Study of 1988 (NELS-88)

(Gerber, 1996). In the open-ended mathematics section, students left an average of 3.06 of 17 subitems blank. This research also examined the effects of gender and ethnicity. On some items, there was almost no gender difference in likelihood of responding to all parts of the item; on other items girls were about 2/3 as likely to complete all parts. African-American and Hispanic students were less than half as likely to complete all parts of an item as nonminority students were. However, when students' scores on the multiple choice section and their perceptions of the difficulty of previous items were controlled, there were no gender or ethnicity effects. The perceived difficulty of previous items (a motivational factor) had a greater influence on minorities than nonminorities.

Again, in all these situations, it is difficult to know whether students know so little about the task they can give no related response or if they simply are unmotivated, especially if there is no external measure of ability. For example, Gerber (1996) found that students who scored high on the multiple choice section were more likely to complete all constructed response items. Part of this could be due to the increased sense of self-efficacy/motivation in high-achieving students, but at least some of it is surely due to increased knowledge about the answer.

Nonresponse is only one measure of motivation. Student responses on a test motivation or attitude scale might be less confounded with performance (though performance can obviously influence these measures as well). In a sample of college students taking a required low-stakes test, those who took a traditional multiple choice test had higher scores on the Student Motivation Questionnaire (Sundre, 1996) than students who took an essay test. The students taking the traditional test had motivation T-scores of 50.6 compared to 46.5 (a difference of about 0.40 standard deviations).

However, the results were confounded by differences in subject areas tested as well as the accompanying tests (the students who wrote the essay wrote in a language arts context and also responded to a multimedia Fine Arts test, while the students who took the traditional multiple choice test responded in the context of natural and social sciences).

Observational reports of task engagement provide another measure of motivation. Constructed response tests involving active manipulation of materials may be more motivating than constructed response tests where all tasks are completed on paper. The Massachusetts Educational Assessment Program used some active mathematics and science tasks in 1989 (Badger, Thomas, & McCormack, 1990). For these tasks, students worked in pairs to use materials to solve problems. Observers reported about 80-95% (depending on the task) of the eighth graders were engaged throughout the task. About 70-85% of the students seemed enthusiastic; almost all others were rated neutral, not low, on enthusiasm. Enthusiasm was lowest for the least-structured tasks. On one task, observers also recorded how carefully the students worked and how concerned they were with the accuracy of their response. Though most students were actively engaged, 23% appeared to have no concern for accuracy. Interesting tasks may motivate students to participate in a test, but this is not always the same as motivating them to do well.

Test-takers' perceptions of validity are likely to be related to motivation; students may try harder when the task seems meaningful. Chan and Schmitt (1997) reported that Black students rated the face validity of a videotaped situational judgment test to be higher than a written version, while White students saw little difference between the face validities of the two formats. Adding face validity to the model for test performance decreased the race by method interaction (without controlling face validity or the reading

skills by test method interaction, the race differences were much greater under the written format).

Attributions and expectancies are associated with motivation and performance (Curran & Harich, 1993; Gerber, 1996; Multon, Brown, & Lent, 1991; Uguroglu & Walberg, 1979; Wolf, Smith, & DiPaolo, 1996). Attributions may be differentially associated with performance depending on response format. Chandler and Spies (1981) found that expectancies of success were more highly correlated with attributions of ability on objective tests than they were on essay tests (students with higher expectancies were more likely than students with low expectancies to think their performance was due to their own abilities on multiple choice tests). Correlations of expectancies with attributions of mood, luck, and help from others were higher on essay tests than objective tests (though they were still lower than correlations of expectancies with ability on both formats). Apparently, students who expected to do well on multiple choice tests were more likely than students who did not expect to do well to think that their performance was due to their own abilities, especially on multiple choice tests. On essay tests, the relationship between expectancy and ability was lowered and the relationship between expectancy and mood, luck, and external help was increased. Students who are confident in their abilities, then, may be more motivated on multiple choice tests, where they believe there is a clearer relationship between performance and ability.

One aspect differentiating multiple choice and constructed response items may be the cognitive level of the items. Though both response formats can address all cognitive levels (from basic factual knowledge to complex synthesis and evaluation), proponents of constructed response items tend to claim they are more appropriate for testing higher-level

cognitive skills. In studies of constructed response and multiple choice items written to assess the same cognitive level, Crooks' (1988) review of the literature found few performance differences. Further, in studies where students were told the response format of the exam but were not given concrete examples of items, there were only small differences in test performance and study behavior. It was only when the students were provided with concrete examples of the types of items they would encounter that there were differences in performance. This summary suggests that it is the students' study behavior in anticipation for the cognitive level of the items, rather than the items themselves, that influences performance differences by response format. However, studying implies high-stakes tests of limited content, primarily classroom tests.

Another difference characterizing constructed response items is that they may be perceived by students as more "work" than multiple choice items. Even if they require the same thought processes, constructed response items require students to physically write the answer (instead of filling-in a bubble) and they often require students to choose words to explain their reasoning. Within the multiple-choice format, Wolf, Smith, and Birnbaum (1995) showed that motivation is affected by the degree to which an item is perceived as "mentally taxing". Educators rated items based on how much "mental energy" they would require from students apart from the "difficulty" (in terms of how many students might get the answer correctly). For example, a long division problem would not be difficult for high school students, but it would be mentally taxing. One group of students taking this test faced consequences (10th graders who would be targeted for remedial work if their scores were too low) and another group did not (11th graders who were taking the test essentially as a pilot/norming test because it would be used as a graduation exam for

future cohorts). A DIF index (comparing the 10th and 11th graders) was computed for each item. This index was correlated with difficulty and taxation; the more taxing or less difficult (meaning the item had a higher p value) the item, the greater the DIF favoring 10th graders (the correlations were -.39 with difficulty and -.55 with taxation, once a single outlier was omitted). This suggests that students in high stakes conditions are more likely to work carefully on easy items (where students in low stakes conditions could be bored and careless) and on mentally taxing items (where low stakes students might not want to extend the necessary effort).

If constructed response items are perceived as more taxing (and even if they require no more effort to solve the problem, they would seem to require more effort to write the answer and explanation), the situation may be similar for constructed response compared to multiple choice items. Under low stakes conditions students may not try as hard on the constructed response items, making them appear differentially harder under low stakes.

Effects of Anxiety on Performance

If motivation is the primary difference between performance on high stakes and low stakes test, it would seem that scores would improve on high stakes tests. However, some have suggested that the relationship between arousal and performance is an 'inverted-u'; performance is highest at moderate level of arousal. Arousal may be motivating up to a point, where it becomes anxiety-inducing. The Yerkes-Dodson Law describes this relationship: "There exists an optimal level of arousal for performance of any given task. Levels of arousal above and below this optimal level will be associated with relatively lower performance" (Smith & Smoll, 1990, p. 437). However, others

contend this relationship may be due to measures of arousal which mix the constructs of arousal and stress. Using instruments designed to measure these constructs separately, Stanley, King, and Glass (1989) found a small negative correlation between stress and mathematics test performance and a small positive correlation between arousal and test performance, with no evidence of a quadratic trend between arousal and performance.

Many have focused not on the more general construct of arousal but on the specific construct of anxiety (the component of arousal hypothesized to decrease performance). In a meta-analysis, Hembree (1988) concluded that anxiety scores were higher in high-stress, ego-involving testing situations, which would include high stakes exams. The correlation between anxiety and test scores tends to be greater (more negative) in studies of standardized tests than in studies of classroom tests (Crooks, 1988); possibly other factors (motivation, preparation) play a bigger role and override some of the effects of anxiety on classroom tests.

In general, the test scores of students who have higher anxiety scores are lower (Crooks, 1988; Hembree, 1988; Wolf & Smith, 1995; Wolf, Smith, & DiPaolo, 1996). This could be because anxiety interferes with cognitive processing during the test or because anxiety interferes with learning (then the low test scores would be an accurate reflection of how much the students learned). Another possibility is that anxiety is caused by the students' awareness of lack of ability. Tobias (1985), after reviewing the literature, concluded at least some part of the relationship between anxiety and test scores was due to the effects of anxiety during the test, because the relationship between anxiety and performance is stronger under "ego-involving" testing situations.

More evidence that anxiety interferes with test-taking is shown in the work of Hill (1980) and Plass and Hill (1986). In these studies, conditions were manipulated to reduce test anxiety When time limits for completing a standardized test were relaxed, the test scores of middle school students who had moderate or high levels of anxiety (with anxiety measured as a trait, not specific to a given test administration) increased compared to the standard timed condition (Hill, 1980). When students were warned that some items would be difficult and they should not worry about missing some items, scores did not improve for any group and the scores of the low anxious children decreased. These instructions might have decreased motivation as much as they decreased anxiety. The most highly anxious group, however, did have higher scores under a combined more time, lower expectation condition. Plass and Hill (1986) included gender in the design of a similar study of the effects of timing/not timing third and fourth graders on a standardized mathematics test. The highly and moderately anxious boys did better when they were not timed; their scores in this condition were not significantly different from the scores of the low-anxious boys. Low anxious boys and high anxious girls, however, did somewhat better when they were timed. The authors speculated that the highly anxious girls may have disliked taking mathematics tests so much that a test of unlimited duration was even more aversive than a test they were unable to complete in the allotted time.

Hembree (1988) concluded in a meta-analysis that Hispanic high school students had higher anxiety levels than Black or White high school students. Results varied among younger students; by upper elementary school Hispanic students had higher levels of anxiety, and at some grade levels Blacks were more anxious than Whites. In Hill's (1980) sample of fourth through eighth graders, Hispanic students had higher levels of test

anxiety than Black students, who had higher levels than White students. Not only did the levels of anxiety vary by ethnicity, but the correlation between anxiety and performance differed somewhat as well. This correlation was -.36 for Whites, -.41 for Blacks, and -.45 and -.51 for Hispanic students in two types of bilingual or ESL classes (the correlation was even more negative, -.93, for Hispanic students completely in mainstream classes, but there were only eight students in this group).

Studies which have manipulated the consequences of the exam have shown no evidence of a moderating effect of consequences on the relationship between anxiety and performance. In Wolf and Smith's (1995) study of students who took one form of a test which would affect their course grade and another form which would not, anxiety was equally associated with test scores regardless of the consequences (correlations of -.28 and -.29). In this study, anxiety was treated as a stable characteristic, measured five days prior to the tests, rather than an effect of the test consequences. Wolf, Smith, and DiPaulo (1996) found that students were more anxious in the consequential condition (the groups differed by 0.46 standard deviations), but anxiety had a significant effect on performance under both conditions.

Though anxiety has been found to be associated with lower test scores, and higher consequences would seem to provoke some anxiety, performance has generally been shown to increase with higher consequences, on average (see "Effect of Consequences on Test Scores", this chapter). For most students, the motivation of higher stakes seems to override any increased anxiety. Anxiety is not measured in the present study; the literature on anxiety was briefly described here to acknowledge the possible negative effects of anxiety (logically linked to test consequences) on the scores of some students.

Summary of the Literature

Average scores are generally higher on high stakes tests (Burke, 1991; Taylor & White, 1981; Wolf & Smith, 1995; Wolf, Smith, & DiPaulo, 1996). One reason for this could be motivation. Motivation tends to be low on low stakes tests such as the NAEP and standardized tests with few individual stakes (Kiplinger & Linn, 1992; Paris, Lawton, & Turner, 1992; Paris, Turner, Lawton, & Roth, 1991), and it is lower under non-consequential conditions than consequential conditions (Arvey, Strickland, Drauden, & Martin, 1990; Wolf & Smith, 1995; Wolf, Smith, & Birnbaum, 1995; Wolf, Smith, & DiPaulo, 1996). If "pattern-marking" (marking bubbles in a set pattern) is indicative of very low motivation, extremely low motivation is more common among boys than girls and among minorities than non-minorities on low stakes tests (Freund & Rock, 1992).

In turn, motivation is positively associated with performance (Multon, Brown, & Lent, 1991; Uguroglu & Walberg, 1979), though one study of adult job applicants found this association only for one of three tests (Arvey, Strickland, Drauden, & Martin, 1990). Some have found the relationship to be higher for boys than for girls (Karmos & Karmos, 1984), but Brown and Walberg (1993) found no interaction between motivation and gender, and across many studies Uguroglo and Walberg (1979) found no average differences in the correlation by gender.

Interest may be a factor in motivation (Arvey, Strickland, Drauden, & Martin, 1990), and interest increases reading comprehension scores (Asher, 1979; Asher, 1980; Bernstein, 1955; Stevens, 1979). There is no evidence that the association between interest and test scores varies by ethnicity (Asher, 1980). The findings of the effects of gender on the correlation are mixed, with some studies showing a stronger relationship for

boys (Asher, 1980; Bernstein, 1955) and others showing no gender differences (Asher, 1979; Stevens, 1979).

The response format of items may influence the degree of group differences on a test. Results are mixed, with most studies finding that adverse impact related to minority group status seems to be as large or larger on constructed response items as on multiple choice items (Bond, 1995; Feinberg, 1990; Linn, Baker, & Dunbar, 1991), though on one series of tests (Badger, 1995) the relative position of minorities increased on the constructed response items (minority students scored lower on both item types, but the difference was smaller on constructed response items). One reason for group differences on constructed response items is that the smaller number of independent tasks on constructed response items, due to the longer time involved and the often greater contextualization provided, allow for a smaller variety of topics (Feinberg, 1990; Linn, Baker, & Dunbar, 1991; Messick, 1994). Group differences in interests or experiences have less chance to balance out when there are fewer topics.

If some of the adverse impact on low stakes tests is due to especially low motivation of minority students (suggested by Freund and Rock's (1992) findings of greater "pattern-marking" by minorities), and if motivation on high stakes tests is more uniform across ethnic groups and response formats, it would be expected that performance would increase more for groups (and items) for which motivation was lower on the low consequence test (the smaller motivation gap under high stakes would lead to a smaller performance gap). If a topic held less interest for one group (but the students had adequate background knowledge to respond appropriately when motivated), that group would be expected to show greater increases in performance when the stakes were

increased. On the other hand, if most of the adverse impact were due to lower topic knowledge, greater motivation would not have much impact on the scores of the group with low topic knowledge. The curricula delivered in disadvantaged schools tends to be of lower quality and quantity (Badger, 1995; Darling-Hammond, 1995; Dreeben & Gamoran, 1986), so disadvantaged students may have lower topic knowledge and experience.

Motivation and performance may be influenced by item response format. Nonresponse is one indication of low motivation, and nonresponse tends to be high on low stakes constructed response items (Badger, 1989; Freund & Rock, 1992; Gerber, 1996; Wainer, 1993). In one sample, nonresponse on constructed response items was higher among minority students, though there was no ethnicity effect when performance on multiple choice items was controlled (Gerber, 1996). Students taking a low stakes essay test had lower motivation than students taking a multiple choice test (Sundre, 1996). Constructed response items may be perceived as more "work", and Wolf, Smith, and Birnbaum (1995) showed that scores on "mentally taxing items" were affected more by the consequences of the test.

Anxiety is negatively correlated with test scores (Crooks, 1988; Hembree, 1988; Wolf & Smith, 1995; Wolf, Smith, & DiPaolo, 1996), and high stakes tests increase anxiety (Hembree, 1988; Wolf, Smith, & DiPaulo, 1996). Highly anxious students generally perform better under less stressful conditions (Hill, 1980; Plass & Hill, 1986). Hispanic students have greater test anxiety (Hembree, 1988; Hill, 1980), and the relationship between anxiety and performance is at least as strong (somewhat stronger) for minorities (Hill, 1980). The consequences of the test, though, do not seem to change the

relationship between anxiety and test scores (Wolf & Smith, 1995; Wolf, Smith, & DiPaulo, 1996). Though anxiety levels are higher on high stakes tests, given the findings described earlier on test consequences and performance, anxiety apparently does not increase enough to offset the positive effects of increased consequences.

As noted, the present study did not measure motivation or anxiety. Rather, these constructs were described as a possible explanation of why test performance increases on tests which have greater consequences, and why students might be less likely to leave items blank on a test with higher consequences. The literature summarized above did not specifically study how item response format affects the degree of change in performance. Most of the work on test consequences used only multiple choice items (Burke, 1991; Freund & Rock, 1992; Taylor & White, 1981, Wolf & Smith, 1995; Wolf, Smith, & Birnbaum, 1995; Wolf, Smith, & DiPaulo, 1996). Some researchers observed that motivation was particularly low on constructed response items without presenting evidence that this had a larger performance impact for constructed response items than for multiple choice items (Badger, 1989; Badger & Thomas, 1989a; Badger & Thomas, 1989b; Freund & Rock, 1992; Thomas, 1989). While Freund and Rock (1992) showed that Blacks and males were particularly unmotivated on a low-stakes test, they did not test students under more consequential conditions to see if motivation (or performance) changed differently for Blacks and males compared to Whites and females.

In the present study, the following questions were addressed:

(1) What effects do test consequences, gender, and ethnicity have on response rate? Response rate is one indication of motivation, and results related to this question contributed information about whether or not the listed factors were associated with

motivation. If the results were similar to the results related to test scores (see question 2), then motivation would be a likely explanation for the findings about test performance.

(2) What effects do test consequences, response format, gender, and ethnicity have on test scores?

As noted, previous literature showed that higher consequences increased test scores. Results from this question added information about possible interactions between consequences and the other factors.

(3) How well do item parameter estimates based on pilot data fit the operational data, and how do pilot estimates based on treating omitted items as not-administered compare to estimates based on treating omitted items as incorrect?

Significant interactions found for the first two questions would lead to differences in item fit. Results from this question were used to assess how much of an impact differences in response rate and test performance had on item fit.

CHAPTER 3

METHOD

Participants

Students participated in either the spring-semester 1996 or spring-semester 1997 Michigan High School Proficiency Test (HSPT), or the piloting of the test forms later administered at these testing times. For science, both the 1996 and 1997 tests counted towards diploma endorsements; the first year (1996) is reported in detail, and the second year (1997) is described in the Appendix. All information in Chapters 3 and 4 concerns the 1996 form unless reference is specifically made to the second science cohort. In math, the 1996 test did not affect the diploma endorsement for most students, so only the 1997 form was analyzed.

In science, in the non-consequential (low stakes) test condition there were 512 White females, 89 Black females, 487 White males, and 59 Black males. In math, there were 579 White females, 69 Black females, 548 White males, and 62 Black males. Blacks were the only ethnic group used in these analyses because there were insufficient numbers of other minority groups. The students in the science sample attended 28 different schools; approximately 1/3 of the 11th grade students in each of these schools took this form of the pilot test (a 29th school was originally in the pilot sample, but data from this school was discarded because none of the high-stakes students in this school indicated their ethnicity). In math, the students attended 32 schools. Schools were randomly

selected to participate (from all population/urbanicity strata within the state), and students within schools were randomly assigned to test forms (two additional forms, to be used in other testing periods, were administered to the other 11th graders in these schools).

Participants in the consequential (high stakes) test condition were students who were tested during the regular spring-semester administration of the HSPT. Only the results from the schools which participated in the pilot-testing of this test form were used in this analysis. In these schools, all students who indicated their ethnic/racial group was either "White" or "Black" were selected for analysis. In science there were 198 Black females, 1272 White females, 149 Black males, and 1255 White males. In math there were 206 Black females, 935 White females, 148 Black males, and 954 White males.

The students in the selected schools were similar to the state population of 11th graders, in terms of their average scores on the high stakes test and their ethnic background. In the population of 11th graders tested at the 1996 administration (science), about 22% of those who indicated their ethnicity identified themselves as non-white, compared to 21% in the selected schools (though 26% of the students did not respond to this question). In the population of 11th graders tested at the 1997 administration (math), about 25% of those who indicated their ethnicity identified themselves as non-white, compared to 22% in the selected schools. However, 41% of the students in the selected schools did not identify their ethnicity, compared to 30% of all tested students. In schools where less than 75% of the students indicated their ethnic group, the school proportion minority was based on the pilot data, where there was greater information.

Limiting the sample to Blacks and Whites who indicated their gender, the mean science scale score was 383.7 (SD = 37.9) in the larger group and 385.9 (SD = 37.5) in

the selected schools. In math, the mean scale score was 401.00 (SD = 50.3) in the larger group and 404.04 (SD = 51.4) in the selected schools. Including students of all ethnic groups, students who did not indicate their ethnicity scored *slightly* lower than students who did, but the difference was similar across gender. In math, the girls who supplied ethnic information averaged 2.5 points higher than those who did not; the difference was 3.2 for boys. In science, the difference was 6.2 for girls and 8.9 for boys.

In both groups, students who responded to less than half the multiple choice items were excluded from the analyses. In science, five low-stakes students and one high-stakes student were excluded for this reason and were not included in the numbers above. These students answered several items at the beginning, but then appeared to quit early in the test. In math, 10 low-stakes students and 16 high-stakes students were excluded for this reason. Under high-stakes, all these students responded to items on only one of the two days of testing; under low-stakes, some of these students responded to only a few items.

Instrument

The Michigan High School Proficiency Test (HSPT) had four components during the years studied here: mathematics, science, reading, and writing. The tests were *not* designed as minimum competency exams; they were intended to reflect high school level (through the end of the sophomore year) skills. Beginning with the graduating class of 1997, students who scored in the proficient category received endorsement seals on their diplomas at graduation. Separate endorsements could be earned in mathematics, science, and communication arts. Students initially took the tests in their junior year of high school, and students who did not earn endorsement seals had opportunities for retaking the tests.

The HSPT was first administered as an operational test (leading to diploma endorsements) in the spring of 1996. The science section administered during this testing period counted for diploma endorsement. The mathematics section administered in 1996 was not high-stakes for most students because this cohort had had an opportunity to earn diploma endorsements in mathematics (and reading) the previous year based on a discontinued statewide examination of 10th graders. The mathematics form analyzed here was the form administered in the spring of 1997, the first year the mathematics test was widely used for diploma endorsement.

Each form of the science section of the HSPT consists of 42 multiple choice items and eight constructed response items. Only 34 of the multiple choice items were analyzed here because eight of the items were different on the pilot test (for equating purposes). The maximum point value of the constructed response items varies across forms (a different form is developed for each administration), but on this test form, each constructed response item was worth two points, for a total of 16 points.

Two of the constructed response items are related to a scientific investigation described in the test, and two are responses to a text passage related to science; the other four items involve applications of science knowledge or methods. These items could be considered intermediate on the "performance" continuum; the students do not actually carry out the investigations they design or expand, and they interpret or explain supplied results or described phenomena rather than those they have generated or observed (or the students may be asked to create a pattern of results or a scenario that would support a hypothesis). Some items require more content knowledge and others require more procedural knowledge, as is true for the multiple choice items as well.

The two investigation items were highly related, both conceptually and statistically. The correlation between these two items was .60, compared to correlations of .19-.33 for all other pairs of constructed response items. Therefore, these items were treated as a single item worth four points (nine categories when 1/2 points were considered). Effectively, then, there were seven constructed response items. The two constructed response items based on the same text passage, as well as several clusters of multiple choice items (each cluster was written to relate to a common scenario or graphic) might also be logically viewed as more interdependent than items on the test as a whole. To check for this possibility, Yen's Q3 (Yen, 1993), was calculated for all item pairs, using operational data. For this statistic, an expected score is calculated for each student on each item, based on an item response function (Yen developed the statistic for the three-parameter model, but here the one-parameter model was used, as explained in the analysis section). The residual between each student's observed item score and his/her predicted item score is then found, and these residuals are correlated for pairs of items. Yen (1993) suggested a cutoff of .20 in deciding whether the assumption of local independence was violated to an extent which would make a practical difference. After combining the two investigation items, only two pairs of items met this criteria for the HSPT science test. One pair of multiple choice items had a residual correlation of .23; these items were treated as a set (scored 0-2) in item and ability estimation. Another pair, with one multiple choice and one constructed response item (adjacent and on the same theme), had a residual correlation of .21. These items were left separate because later analyses required separate scores on the multiple choice and constructed response sections. Two other pairs had residual correlations between .10 and .20. One was another mixed-format pair, which

was left as two separate items, and the other was a multiple choice pair, which was combined into a single item (scored 0-2). All other correlations were less than .10.

The math section of the HSPT consists of 40 multiple choice items and 6 constructed response items (worth 2-5 points each). Only 32 of the multiple choice items were analyzed here because 8 items were not consistent from pilot to operational test. As in science, the constructed response items involve written responses, which may include a drawing or figure and usually require words to be answered completely. In these items, students are expected to explain their solutions and reasoning processes.

There was no reason to expect to find sizable residual correlations on the math section, but for consistency with the science analyses the local item independence assumption was checked with Q3. Again, the operational data were used for the item calibration, and all items were estimated together. One pair of constructed response items was on the edge of Yen's (1993) suggested limit of .20, but I did not combine them because there were already so few constructed response items. There were four multiple choice items with intercorrelated residuals: five of the correlations were greater than .20, and the sixth was greater than .10. These items covered three different content areas, and only two were adjacent; there was no conceptual reason to expect them to be related. I summed these four items and treated them as one item with five ordered score categories.

For the pilot test, the students' answers to the constructed response items were read by two raters, and if the scores were more than one point apart an additional rater scored the response. The item score was the average of these two (or three) ratings (if three raters were used, the score was rounded to the nearest .5 for the parameter estimation analyses in this study; rounding was necessary for less than 1% of the pilot

students). In the operational test, if a third rater was needed, a more highly skilled rater was used and the rating of this "expert" replaced the ratings of the other two raters.

Analysis

Response/Non-Response.

For the analysis of response rate, each item was coded one or zero for each student. If the student wrote any answer related to a constructed response item prompt, this response was coded "1". Blanks and completely irrelevant responses (i.e. "this is a stupid question") or "don't know" were coded "0". The combined items were coded "1" if the student responded to either part. These dichotomous codings are problematic for common linear models. The errors are not likely to be normally distributed or homogeneous. Additionally, errors are unlikely to be independent within students, and the errors may be correlated within schools as well. If the responses follow a binomial distribution, a hierarchical generalized linear model which uses the estimated log-odds of response as the outcome variable, as operationalized in the software package HLM 4 (Bryk, Raudenbush, & Congdon, 1996), can deal with these complications.

Response rates for the multiple choice items were very high (over 99%), so only the constructed response items were analyzed. The proposed model had three levels. At the lowest level, student j's (nested within school k) response to item i was a function of the student's average log-odds of response. At the next level (level-2), each student's response propensity was modeled as a function of the school mean and the effects of gender, ethnicity, and stakes (and their interactions) within that school, with an error term for individual student differences. At the highest level, each of the level-2 coefficients was

a function of the grand mean, the proportion of minority students in the school, and random variance.

At the first level, the outcome variable in the linear model was not the dichotomous response but a transformation of the underlying probability of response. The transformation used was the log of the odds of response because the resulting error (unexplained variance) should be homogeneous and normally distributed with this transformation, if the responses follow a binomial distribution within students. Also, this outcome variable has no lower or upper bounds, with a value of zero when the probability of response is 0.5.

$$\text{Level 1: } \log\left(\frac{P(y_{ijk}=1)}{1-P(y_{ijk}=1)}\right) = \pi_{0jk}, \qquad (1)$$

where π_{0jk} is the log of the odds of student j (in school k) responding to any item i. The first subscript (0) signifies this is the intercept--additional within-student coefficients would be labeled π_{1jk}, π_{2jk}, etc.

At the second level, each student's log-odds of response was then predicted by student characteristics:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(X_{1jk}) + \beta_{02k}(X_{2jk}) + \beta_{03k}(X_{3jk}) + \beta_{04k}(X_{1jk}X_{2jk}) + \beta_{05k}(X_{1jk}X_{3jk}) + \beta_{06k}(X_{2jk}X_{3jk}) + \beta_{07k}(X_{1jk}X_{2jk}X_{3jk}) + r_{0jk}, \qquad (2)$$

where $X_{1jk} = 0.5$ for high stakes and -0.5 for low stakes, $X_{2jk} = 0$ for Whites and 1 for Blacks, centered around the proportion of Blacks in the total population (i.e., the code for White = 0 - proportion of Blacks) , $X_{3jk} = 0$ for females, 1 for males, centered around the proportion of males in the high stakes sample. These codings resulted in an intercept (β_{00k}) which was the predicted mean (of the log-odds of response) for school k, with half

the students taking a pilot test and half taking an operational test, adjusted for ethnicity and gender (based on the high-stakes condition, because students were not sampled within schools under this condition, which should lead to more stable estimates). If the gender and ethnic codes were not centered, the intercept would be the predicted score for White females (the group with $X_2 = 0$ and $X_3 = 0$). While there were not equal numbers of students taking the pilot and operational tests, the test-stakes codes were not intended to make the intercept reflect the average test conditions (the number of students in each condition was based on practical reasons, not on some average mix of conditions in the population of tests). The -0.5 and 0.5 codings allowed for an easy interpretation of the coefficients.

The first subscript of each X identifies which factor it is associated with (1 for stakes, 2 for ethnic group, 3 for gender), and the j,k connect it with student j in school k. The first subscript of each β links it to a particular π with the same initial subscript, the second subscript identifies which X it is associated with, and the third subscript (k) stands for school k.

Effects were later removed if they were not significant, except that main effects were not removed if related interactions were significant.

At the third level, school effects were modeled as a function of the proportion of students in the school who identified themselves as non-White. This proportion was based on all students, including those who were not used in the other analyses because they belonged to ethnic groups other than Black or White. In some schools, a large proportion of high-stakes students did not indicate an ethnic group. In the math (1997) sample, non-response to this question was greater than 50% in 13 of the 32 schools. In these schools,

the proportion minority variable was instead based on the pilot data (where, in every school, over 80% of the students indicated their ethnicity). This procedure was also used for six schools on the science test. All other schools had response rates of at least 80% to the ethnicity question.

Theoretically, the model would also have a random effects component for each β; schools vary in how large the effects of stakes and student characteristics are. However, as noted, it was difficult to estimate the variance of these random effects, especially with only 28/32 schools, so a random effects term was initially included only for the school mean. The model was:

Level-3: (3)

$$\beta_{00k} = \gamma_{000} + \gamma_{001}(W_k) + u_{00k}.$$
$$\beta_{01k} = \gamma_{010} + \gamma_{011}(W_k)$$
$$\beta_{02k} = \gamma_{020} + \gamma_{021}(W_k)$$
$$\beta_{03k} = \gamma_{030} + \gamma_{031}(W_k)$$
$$\beta_{04k} = \gamma_{040} + \gamma_{041}(W_k)$$
$$\beta_{05k} = \gamma_{050} + \gamma_{051}(W_k)$$
$$\beta_{06k} = \gamma_{060} + \gamma_{061}(W_k)$$
$$\beta_{07k} = \gamma_{070} + \gamma_{071}(W_k),$$

where the γ's are means, W_k is the proportion of students who are minorities, centered around the weighted mean of the schools in the sample, and the u's are random school effects. The first two subscripts on the γ link it with a particular β, and the third subscript identifies the order in the sequence of γ's predicting a β; 0 is the third subscript on the intercept, 1 is the third subscript on the coefficient for the first predictor, 2 would be the subscript for the next predictor if there were one (and the W's would be correspondingly numbered).

γ_{010} through γ_{070} (the averages of β_{01k} through β_{07k}) represented the effects of gender, stakes, ethnicity, and their interactions. The variance of u_{00k} is the variance in school means, after controlling the proportion of students who were minorities.

Test Scores

The item scores were more problematic than the response/non-response codings. Most of the items were dichotomous right/wrong (the multiple choice items), but some had many score points. Instead of working with individual item scores, two composite scores were predicted for each student, one based on the multiple choice items and one based on the constructed response items. These ability estimates were based on item response theory. The one parameter model was used because that is the model actually used to equate and score the tests. The constructed response items were estimated with the one-parameter partial-credit model, simultaneously with the multiple choice items. Thus, in the calibration sample, the average scores (estimated abilities) were equal for both types of items (though subgroups of students could have systematic differences in the two scores). The high-stakes responses were used to calibrate the items, under the assumption that effort would be more constant (within students, across items) on the operational test, leading to better fit between the model and the data. On the no-stakes test, if some students *and* some items were more influenced than others by lack of effort, poor fit could result (if a student's low effort uniformly depressed his performance on all items, or a certain set of items tended to elicit little effort from all students, this would not be a problem for fit; it is the student by item interactions which lead to poor fit).

For the actual HSPT, item estimates and person estimates are obtained by joint maximum likelihood. For this study, item estimates were instead obtained by marginal

maximum likelihood, using PARSCALE 3.3 (Muraki, E., & Bock, R.D., 1997), with the ability distribution empirically estimated. For long tests (such as the HSPT), correlations between estimates under joint maximum likelihood and marginal maximum likelihood tend to be high and close to the true parameters (Abdel-fattah, 1994; Lord, 1986; Mislevy & Stocking, 1989, Stocking, 1989; Yen, 1987).

After the items were estimated, Bayesian Expected A-Posterior (EAP) scores were estimated for each person on the two subscales, with a normal prior distribution. The scores were also estimated under maximum likelihood (MLE). While the correlation between the MLE and EAP scores was .99, the EAP scores resulted in ability distributions which were more similar for the two subscales than the distributions estimated under maximum likelihood. In the calibration sample, for science the standard deviation using EAP scores was 0.76 for the multiple choice abilities and 0.84 for the constructed response abilities. Using maximum likelihood scores, the standard deviations were 0.91 and 1.27. In math, the standard deviations were 1.06 (multiple choice) and 0.88 (constructed response) using EAP scores, compared to 1.36 and 1.16 using MLE scores. If the underlying ability for the two types of items is the same, the distributions of abilities should be similar. Also, for shorter tests (such as the constructed response scale here), Bayesian scores tend to be closer to the true abilities than maximum likelihood scores, though they are somewhat biased (Abdel-fattah, 1994; Lord, 1986; Yen, 1987). Bayesian estimates tend to be reasonably accurate even when the true distribution departs from normality and the prior distribution is specified as normal (Reise & Yu, 1990; Yen, 1987). Though reported scores on the HSPT are estimated by joint maximum likelihood, scores are reported only for the total set of items, not for the subscales used in this study. These

shorter scales introduced different problems (less information to estimate abilities, and a greater proportion of students with 0 or perfect scores) which made EAP estimates more appropriate.

Two ability scores, then, were nested within each student, and students were nested within schools. The level one model was:

$$\hat{\theta}_{ijk} = \pi_{0jk} + \pi_{1jk}(a_{ijk}) + e_{ijk}, \qquad (4)$$

where $\hat{\theta}_{ijk}$ is the predicted ability score under format i for student j in school k, π_{0jk} is the average score of student j (in school k), a_{1ijk} = -0.5 for multiple choice and 0.5 for constructed response, so π_{1jk} is the difference between the two scores (the format effect). The first subscript for π identifies whether it is the intercept (0) or the coefficient for the first predictor (1)--additional within-student coefficients would be labeled π_{2jk}, π_{3jk}, etc. The measurement error is e_{ijk}, and its variance differed for each ability estimate for each student. This variance and the $\hat{\theta}_{ijk}$ were estimated in the IRT software (PARSCALE 3.3).

The level-two model used the same predictors as the level-two model for response/non-response, and the same predictors were used for both the intercept (π_{0jk}) and the format effect (π_{1jk}).

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(X_{1jk}) + \beta_{02k}(X_{2jk}) + \beta_{03k}(X_{3jk}) + \beta_{04k}(X_{1jk}X_{2jk}) + \beta_{05k}(X_{1jk}X_{3jk}) + \beta_{06k}(X_{2jk}X_{3jk}) + \beta_{07k}(X_{1jk}X_{2jk}X_{3jk}) + r_{0jk}, \qquad (5)$$

$$\pi_{1jk} = \beta_{10k} + \beta_{11k}(X_{1jk}) + \beta_{12k}(X_{2jk}) + \beta_{13k}(X_{1jk}X_{2jk}) + \beta_{14k}(X_{1jk}X_{2jk}) + \beta_{15k}(X_{1jk}X_{3jk}) + \beta_{16k}(X_{2jk}X_{3jk}) + \beta_{17k}(X_{1jk}X_{2jk}X_{3jk}) + r_{1jk}, \qquad (6)$$

where X_{1jk} = 0.5 for high stakes and -0.5 for low stakes, X_{2jk} =0 for Whites and 1 for Blacks, centered around the proportion Black in the high stakes sample, X_{3jk} =0 for females, 1 for males, centered around the proportion of males in the high stakes sample.

The first subscript of each X identifies which factor it is associated with (1 for stakes, 2 for ethnic group, 3 for gender), and the j,k connect it with student j in school k. The first subscript of each β links it to a particular π with the same initial subscript, the second subscript identifies which X it is associated with, and the third subscript (k) stands for school k.

At the third level, school effects were modeled as functions of the proportion of students who were non-White. A random term was included in the model for the school mean ability and school mean format effect (β_{00k} and β_{10k}). The variance of these terms would show the conditional variances (controlling percent minority) in school means.

Level-3: (7)

$$\beta_{00k} = \gamma_{000} + \gamma_{001}(W_k) + u_{00k}, \quad \beta_{10k} = \gamma_{100} + \gamma_{101}(W_{1k}) + u_{10k},$$
$$\beta_{01k} = \gamma_{010} + \gamma_{011}(W_k), \quad \beta_{11k} = \gamma_{110} + \gamma_{111}(W_k),$$
$$\beta_{02k} = \gamma_{020} + \gamma_{021}(W_k), \quad \beta_{12k} = \gamma_{120} + \gamma_{121}(W_k),$$
$$\beta_{03k} = \gamma_{030} + \gamma_{031}(W_k), \quad \beta_{13k} = \gamma_{130} + \gamma_{131}(W_k),$$
$$\beta_{04k} = \gamma_{040} + \gamma_{041}(W_k), \quad \beta_{14k} = \gamma_{140} + \gamma_{141}(W_k),$$
$$\beta_{05k} = \gamma_{050} + \gamma_{051}(W_k), \quad \beta_{15k} = \gamma_{150} + \gamma_{151}(W_k),$$
$$\beta_{06k} = \gamma_{060} + \gamma_{061}(W_k), \quad \beta_{16k} = \gamma_{160} + \gamma_{161}(W_k),$$
$$\beta_{07k} = \gamma_{070} + \gamma_{071}(W_k), \quad \beta_{17k} = \gamma_{170} + \gamma_{171}(W_k),$$

where the γ's are means, W_k is the proportion of students who are minorities, centered around the weighted mean of the schools in the sample, and the u's are random school effects. The first two subscripts on the γ link it with a particular β, and the third subscript identifies the order in the sequence of γ's predicting a β; 0 is the third subscript on the intercept, 1 is the third subscript on the coefficient for the first predictor, 2 would be the subscript for the next predictor if there were one (and the W's would be correspondingly numbered).

Level-2 effects, and their level-3 counterparts, were removed if they were not significant. If an effect was significant for the format effect, it was also left in the model for the mean to keep the interpretation clearer.

Item Parameter Estimates

To see how well the pilot parameter estimates fit the operational data, item parameters were estimated from the pilot data, again using the one-parameter model (partial-credit model for the polytomous items). Initially, omitted items were treated as incorrect, because students had adequate time to attempt all items if they wished. Parameters were estimated separately for each gender. Based on these item estimates, the abilities of the operational students were re-estimated. Then a measure of fit, the standardized appropriateness index of Drasgow, Levine, and Williams (1985) was calculated for each person, using the formula:

$$z_h = \left[l_{0,h} - E_h(\hat{\theta}) \right] \div \sigma_h(\hat{\theta}), \tag{8}$$

where

$$l_{0,h} = \sum_{i=1}^{n} \sum_{j=1}^{A} \delta_j(v_i) \log P_{ij}(\hat{\theta}), \tag{9}$$

$$E_h(\hat{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{A} P_{ij}(\hat{\theta}) \log P_{ij}(\hat{\theta}), \tag{10}$$

and

$$\sigma^2{}_h(\hat{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{A} \sum_{k=1}^{A} P_{ij}(\hat{\theta}) P_{ik}(\hat{\theta}) \log P_{ij}(\hat{\theta}) \log(P_{ij}(\hat{\theta}) / P_{ik}(\hat{\theta})). \tag{11}$$

The total number of items is n, A is the number of categories (including 0 in this case), $\delta_j(v_i) = 1$ if the student scored in category j on item i and 0 otherwise, and $P_{ij}(\hat{\theta})$ (the probability of category j on item i, given $\hat{\theta}$) is calculated from the item parameters.

This index is approximately normally distributed when item parameters are known. Empirical distributions, based on item estimates from samples in which most examinees fit the model, have been found to approximate the normal distribution fairly well in a practical sense (Drasgow, Levine, & Williams, 1985).

For my purposes, the multiple choice items were treated as having only two options, correct and incorrect, because I was not interested in patterns of choice of distractors. Non-response was counted as zero, not as another option, because it was scored that way on the operational test. Students were flagged as misfitting if they had standardized indices less than -2.58, which would be expected for only 0.5% of the sample if these students' response patterns fit the item estimates.

These analyses were then repeated using item parameter estimates obtained when treating omitted items as if they were not-administered rather than incorrect. For the test of fit, the non-responses of the high-stakes were scored as zero; it was only in the item estimation (using the pilot data) that omitted items were treated as not-presented. If non-response was due to different factors under low-stakes than under high-stakes (such as low motivation under low-stakes), not scoring omitted items from the pilot test could produce item estimates which fit the response patterns of the high stakes students better.

These analyses were not computed by ethnic group because, with less than 200 minority students (both genders combined) on the pilot forms, the parameter estimates were unlikely to be stable, especially for the polytomous items.

The standardized appropriateness fit index flags students whose responses do not fit the pattern expected from the item difficulties. Another way of looking at the fit of items and persons is to flag items on which many people seem to have unexpected

responses, given ability and item difficulty. To do this, the OUTFIT statistic (Wright & Masters, 1982) was calculated based on the differences between observed and expected scores. For item i,

$$OUTFIT_i = \Sigma z^2_{ni}/N, \qquad (12)$$

where N is the total number of persons,

$$z_{ni} = \frac{x_{ni} - E_{ni}}{\sqrt{\sum_{k=0}^{m}(k - E_{ni})^2 \pi_{ni}}},$$

x_{ni} is the observed score for person n on item i,

k is the score point (range 0-m),

π_{nik} is the probability of person n scoring k on item i, and

E_{ni} (the expected score for person n on item i) $= \sum_{k=0}^{m} k\pi_{nik}$.

OUTFIT is more sensitive to unexpected outliers then a related measure, INFIT, in which each residual is weighted by its variance. Someone who has a high probability of a high score on an item but earns a low score instead (or someone who has a low probability of a high score but nevertheless obtains one) will have a larger impact on the OUTFIT statistic than on the INFIT statistic.

The expected value for OUTFIT, if the responses fit the model, is 1. Values less than 1 indicate the responses fit better than expected (less random variance than usual), and values greater than 1 indicate the response patterns do not fit the model as well as expected.

The average OUTFIT was calculated for each gender group by estimation condition (method of treating omits) by response format (constructed response or multiple

choice), for a total of eight averages. These means were compared with a repeated-measures ANOVA, with item as the unit of analysis. In addition to these means, individual items were examined for especially large OUTFIT values (1.5 or more--50% greater than average).

CHAPTER 4

RESULTS

As noted at the beginning of Chapter 3, the science results here pertain to the first year (1996) of the HSPT, and results for the second year are in the Appendix. The math results are provided only for the second year, because the math test did not affect diploma endorsement in the first year.

Response Rate

Response rates for the multiple choice items were very high, averaging over 99% in science (99.6% under low stakes, 99.8% under high stakes) and over 97% in math (97.8% under low stakes, 99.7% under high stakes). Therefore, response/non-response was analyzed only for the constructed response items. Items were nested within students who were nested within schools. Initially, a hierarchical linear model with no predictors was run to partition the variance between-schools compared to between-students within-schools. The outcome variable was the predicted log-odds of response, with the error variance for each student based on the binomial model, given the student's predicted probability of response.

In science, school means could be estimated more reliably than student means (reliability = .935 for schools and .599 for individuals). The estimated average log-odds of response was 2.61, which is equivalent to a probability of 93% (log (93/7) = 2.61). About

24% of the between-student variance was due to between-school differences, leaving 76% due to differences within schools.

In math, differences were also more reliable for school means than for individual students (reliability = .893 for schools and .595 for individuals). The average student had an 88% probability of responding (log-odds = log(88/12) = 1.95). About 18% of the between-student variance was due to between-school differences, and 82% was due to differences within schools.

Next, predictors were added to the model. At the student level, test stakes, gender, ethnicity, and the interactions of these variables were viewed as potential predictors. Stakes were coded -0.5 for low stakes and 0.5 for high stakes. Ethnicity and gender were first coded 0/1 (0 for White, 1 for Black, 0 for female, 1 for male) and then centered around their means in the high-stakes schools, so that the intercepts in the model would be the predicted results for a school with an average number of males and Blacks (because the *average* codes for gender and ethnicity would be 0 in such a school). In the high-stakes group, 49% of the students were male, so after centering, males were coded 0.51 and females were coded -0.49. In science, 12.1% of the students were Black so ethnicity was coded 0.879 for Black and -0.121 for Whites. In math (a new cohort of students), 15.8% of the students were Black, so ethnicity was coded 0.842 for Blacks and –0.158 for Whites. The only effect which was allowed to vary across schools was the school mean; other effects were held constant across schools for ease in estimation. The results appear in Tables 1 and 2.

The model estimated was:

Level-1 Model

$$\log\left(\frac{P(y_{ijk}=1)}{1-P(y_{ijk}=1)}\right) = \pi_{0jk},$$

where y_{ijk} is the observed response of student j (in school k) to item i (1 if an active, relevant response is made, 0 otherwise). The first subscript (0) on π_{0jk} signifies this is the intercept--additional within-student coefficients would be labeled π_{1jk}, π_{2jk}, etc.

Level-2 Model

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(X_{1jk}\text{:Stakes}) + \beta_{02k}(X_{2jk}\text{:Ethnicity}) + \beta_{03k}(X_{3jk}\text{:Gender}) + \beta_{04k}(X_{1jk}X_{2jk}) + \beta_{05k}(X_{1jk}X_{3jk}) + \beta_{06k}(X_{2jk}X_{3jk}) + \beta_{07k}(X_{1jk}X_{2jk}X_{3jk}) + r_{0jk},$$

The first subscript of each X identifies which factor it is associated with (1 for stakes, 2 for ethnic group, 3 for gender), and the j,k connect it with student j in school k. The first subscript of each β links it to a particular π with the same initial subscript, the second subscript identifies which X it is associated with, and the third subscript (k) stands for school k.

Level-3 Model

$$\beta_{00k} = \gamma_{000} + u_{00k},$$
$$\beta_{01k} = \gamma_{010}$$
$$\beta_{02k} = \gamma_{020}$$
$$\beta_{03k} = \gamma_{030}$$
$$\beta_{04k} = \gamma_{040}$$
$$\beta_{05k} = \gamma_{050}$$
$$\beta_{06k} = \gamma_{060}$$
$$\beta_{07k} = \gamma_{070}$$

The first two subscripts on the γ link it with a particular β, and the third subscript identifies the order in the sequence of γ's predicting a β; 0 is the third subscript on the intercept, 1 would be the third subscript on the coefficient for the first predictor (if there were one).

In the tables below and throughout this chapter, the level-3 coefficients are listed below the level-2 coefficients with which they are associated, and the level-2 coefficients are listed below the level-1 coefficients with which they are associated. Only the level-3 coefficients are actually estimated; they represent the average effects across students and schools. In this model, all the level-2 effects are predictors of each student's average (intercept), and each school's level-2 effect is predicted by the mean (intercept) across schools for that effect. The random school effects for the grand mean (u_{00k}) are not actually estimated, though their variance is. The coefficients reported in the tables can be substituted for the γ's in the model, which can be used to estimate the *average* β's, and the β's can then be multiplied by appropriate X's (the X values given in the description of the codes for stakes, ethnicity, and gender) to obtain estimates for a particular ethnic or gender group taking the test under high (or low) stakes. These group estimates are reported later after a final model is chosen.

Controlling student characteristics did not change the within-school variance in science, reducing it only 4%. In math, however, these student characteristics accounted for about 13% of the within-school variance. Tables 1 and 2 show that gender, test stakes, and ethnicity were all significantly associated with the log-odds of response ($p < .01$ for all three effects, γ_{010}, γ_{020}, and γ_{030}, in both subject areas). On average, girls responded more often than boys (γ_{030} was negative and X_3 was positive for males), students in high stakes conditions were more likely to respond than students in low-stakes conditions (γ_{010} was positive and X_1 was positive for high stakes), and Whites were more likely to respond than Blacks (γ_{020} was negative and X_2 was positive for Blacks). In science, there was an interaction between stakes and ethnicity ($\gamma_{040} = 0.510$, $p = .008$);

high stakes produced more of a change in the response rate of Black students. In math, there was an interaction between stakes and gender (γ_{050} =-0.335, p = .001); high stakes resulted in greater changes in the log-odds of response for girls. However, because of the non-linear relationship between odds and probabilities, the probability of response increased (with high stakes) slightly more for boys. These probabilities are shown later in Tables 7 and 8, for the final model.

Table 1 - Student-Level Predictors for Log-Odds of Response, Science

	Coefficient	Standard Error	Approx. T-ratio	df	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	2.711094	0.154924	17.499	27	0.000
For STAKES (S), β_{01k}					
INTERCEPT3, γ_{010}	0.652446	0.142266	4.586	3992	0.000
For ETHNICITY (E), β_{02k}					
INTERCEPT3, γ_{020}	-0.746545	0.194068	-3.847	3992	0.000
For GENDER (G), β_{03k}					
INTERCEPT3, γ_{030}	-0.523382	0.083474	-6.270	3992	0.000
For S X E, β_{04k}					
INTERCEPT3, γ_{040}	0.510278	0.191364	2.667	3992	0.008
For S X G, β_{05k}					
INTERCEPT3, γ_{050}	-0.092514	0.180280	-0.513	3992	0.607
For E X G, β_{06k}					
INTERCEPT3, γ_{060}	0.231921	0.164135	1.413	3992	0.158
For S X G X E, β_{07k}					
INTERCEPT3, γ_{070}	0.175609	0.240172	0.731	3992	0.465

Table 2 - Student-Level Predictors for Log-Odds of Response, Math

	Coefficient	Standard Error	Approx. T-ratio	d.f.	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	2.008863	0.073248	27.426	31	0.000
For STAKES (S), β_{01k}					
INTERCEPT3, γ_{010}	1.335928	0.154946	8.622	3468	0.000
For ETHNICITY (E), β_{02k}					
INTERCEPT3, γ_{020}	-0.811124	0.257083	-3.155	3468	0.002
For GENDER (G), β_{03k}					
INTERCEPT3, γ_{030}	-0.468083	0.060657	-7.717	3468	0.000
For S X E, β_{04k}					
INTERCEPT3, γ_{040}	-0.139183	0.353218	-0.394	3468	0.693
For S X G, β_{05k}					
INTERCEPT3, γ_{050}	-0.335268	0.093038	-3.604	3468	0.001
For E X G, B_{06k}					
INTERCEPT3, γ_{060}	-0.059533	0.132146	-0.451	3468	0.652
For S X G X E, β_{07k}					
INTERCEPT3, γ_{070}	0.171011	0.274219	0.624	3468	0.533

For the next models, a school-level predictor was added. The proportion of students who identified themselves as non-White was used as a predictor of both the intercept (school mean) and the effects of the student characteristics. This proportion was centered around the grand mean. With this centering, school effects would be the predicted effects for a school with an average balance of ethnic groups (after centering, proportion-minority would be zero), with half the students taking a low-stakes test and half taking a high-stakes test.

The first model estimated with this school-level predictor used no student-level predictors, to see how much of the between-school variance in mean response rate could be accounted for by the linear effect of the proportion of minority students. With this term added, the between-school variance was reduced by 44% in science and 45% in math.

For the next model, all the student level factors were included, and proportion-minority (W_k) was added to predict each level-2 effect.

The model estimated was:

Level-1 Model

$$\log\left(\frac{P(y_{ijk}=1)}{1-P(y_{ijk}=1)}\right) = \pi_{0jk},$$

Level-2 Model

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(X_{1jk}\text{:Stakes}) + \beta_{02k}(X_{2jk}\text{:Ethnicity}) + \beta_{03k}(X_{3jk}\text{:Gender}) + \beta_{04k}(X_{1jk}X_{2jk}) + \beta_{05k}(X_{1jk}X_{3jk}) + \beta_{06k}(X_{2jk}X_{3jk}) + \beta_{07k}(X_{1jk}X_{2jk}X_{3jk}) + r_{0jk},$$

Level-3 Model

$$\beta_{00k} = \gamma_{000} + \gamma_{001}\,(W_k\text{:Minority}) + u_{00k},$$
$$\beta_{01k} = \gamma_{010} + \gamma_{011}\,(W_k)$$
$$\beta_{02k} = \gamma_{020} + \gamma_{021}\,(W_k)$$
$$\beta_{03k} = \gamma_{030} + \gamma_{031}\,(W_k)$$
$$\beta_{04k} = \gamma_{040} + \gamma_{041}\,(W_k)$$
$$\beta_{05k} = \gamma_{050} + \gamma_{051}\,(W_k)$$
$$\beta_{06k} = \gamma_{060} + \gamma_{061}\,(W_k)$$
$$\beta_{07k} = \gamma_{070} + \gamma_{071}\,(W_k)$$

If there were additional school-level predictors, the W's would be numbered W_{1k}, W_{2k}, etc., to correspond with the third subscript of the associated γ. With this many school predictors and so few schools, the estimates are not very precise; non-significant effects were eliminated so more precise estimates could be obtained for the remaining effects.

The estimated coefficients for this model are displayed in Tables 3 and 4. Again, the γ's are listed under the β's they predict, but now the predicted β depends on the school's ethnic composition (W_k). Because the proportion-minority is centered around the mean (so an average school has a $W = 0$), the γ's which represent the intercepts are still the averages of their respective β's; a test of one of these intercepts is a test of the average of the effect it represents. For example, the test of γ_{010} is a test of whether the average

stakes effect is significantly different from zero. γ_{001} represents the main effect of school ethnic composition (the effect of proportion-minority on the intercept for school k, β_{00k}). The other γ's which are coefficients to W's (third subscript = 1: γ_{011}, γ_{021}, γ_{031}, etc.) effectively serve as interaction effects because they show how school ethnic composition moderates the stakes effect (γ_{011}), the ethnicity effect (γ_{021}), etc.

Table 3 - Log-Odds of Response on Science CR Items, The Full Model

	Coefficient	Standard Error	Approx. T-ratio	d.f	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	2.749266	0.144220	19.063	26	0.000
MINORITY, γ_{001}	-1.885434	0.744067	-2.534	26	0.018
For STAKES (S), β_{01k}					
INTERCEPT3, γ_{010}	0.875167	0.161280	5.426	3991	0.000
MINORITY, γ_{011}	2.965402	1.068395	2.776	3991	0.006
For ETHNICITY (E), β_{02k}					
INTERCEPT3, γ_{020}	-0.600929	0.311481	-1.929	3991	0.053
MINORITY, γ_{021}	0.591835	0.775899	0.763	3991	0.446
For GENDER (G), β_{03k}					
INTERCEPT3, γ_{030}	-0.473463	0.108083	-4.381	3991	0.000
MINORITY, γ_{031}	1.218082	0.648573	1.878	3991	0.060
For S x E, β_{04k}					
INTERCEPT3, γ_{040}	-0.223209	0.571132	-0.391	3991	0.696
MINORITY, γ_{041}	-2.535735	1.291677	-1.963	3991	0.049
For S X G, β_{05k}					
INTERCEPT3, γ_{050}	-0.399672	0.276334	-1.446	3991	0.148
MINORITY, γ_{051}	-2.462978	1.172343	-2.101	3991	0.035
For E X G, β_{06k}					
INTERCEPT3, γ_{060}	-0.159644	0.403909	-0.395	3991	0.692
MINORITY, γ_{061}	-0.852668	0.849911	-1.003	3991	0.316
For S X GX E, β_{07k}					
INTERCEPT3, γ_{070}	-0.202866	1.030249	-0.197	3991	0.844
MINORITY, γ_{071}	3.962828	2.174412	1.822	3991	0.068

Table 4 - Log-Odds of Response on Math CR Items, The Full Model

	Coefficient	Standard Error	Approx. T-ratio	d.f	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	2.006821	0.102156	19.645	30	0.000
MINORITY, γ_{001}	-1.350982	0.705900	-1.914	30	0.065
For STAKES (S), β_{01k}					
INTERCEPT3, γ_{010}	1.408136	0.095899	14.683	3467	0.000
MINORITY, γ_{011}	0.576021	1.592548	0.362	3467	0.717
For ETHNICITY (E), β_{02k}					
INTERCEPT3, γ_{020}	-0.384300	0.220630	-1.742	3467	0.081
MINORITY, γ_{021}	0.206182	0.828233	0.249	3467	0.803
For GENDER (G), β_{03k}					
INTERCEPT3, γ_{030}	-0.399634	0.102136	-3.913	3467	0.000
MINORITY, γ_{031}	1.171521	0.899497	1.302	3467	0.193
For S x E, β_{04k}					
INTERCEPT3, γ_{040}	-0.243580	0.347431	-0.701	3467	0.483
MINORITY, γ_{041}	-0.521962	1.700159	-0.307	3467	0.759
For S X G, β_{05k}					
INTERCEPT3, γ_{050}	-0.554819	0.148069	-3.747	3467	0.000
MINORITY, γ_{051}	-1.966347	1.032644	-1.904	3467	0.056
For E X G, β_{06k}					
INTERCEPT3, γ_{060}	-0.533490	0.387374	-1.377	3467	0.169
MINORITY, γ_{061}	-.893796	0.974605	-0.917	3467	0.359
For S X GX E, β_{07k}					
INTERCEPT3, γ_{070}	0.514745	0.841908	0.611	3467	0.541
MINORITY, γ_{071}	2.235412	1.478214	1.512	3467	0.130

In science, the within-school stakes by ethnicity interaction was no longer significant when the proportion of minority students was included in the model (γ_{040} = -0.223, p=.696). In high-minority schools, the students were more influenced by the test stakes (γ_{011} = 2.97, p = .006), but when this school effect was controlled there was not an interaction between stakes and ethnicity (though within high-minority schools, high stakes did not increase the responses of Blacks as much as Whites (γ_{041})). The interactions

among the other student-level factors were not statistically significant (as before) and I removed them from the next model. In both this group and the second cohort (see Appendix), the effect of school proportion-minority on the stakes by gender by ethnicity interaction (γ_{071}) was nearly statistically significant (because school proportion-minority modified the stakes by gender by ethnicity effect, there was essentially a four-way interaction), but the effect was different for the two test forms. In the first, in high-minority schools, Black boys had an especially large increase in response under high stakes. In the second cohort, this effect was observed in the low-minority schools. To be able to make some more general interpretations, the main effects model was estimated. The proportion of minority students in the school had a statistically significant effect on the school mean and on the stakes and gender effects; I left the proportion-minority term in the model for these terms, but removed it as a predictor of the ethnicity effect.

In math (see Table 4), the proportion of minority students had a "borderline" significant effect on the intercept ($\gamma_{001} = -1.35$, $p = .065$) and on the gender by stakes interaction ($\gamma_{051} = -1.97$, $p = .056$). Schools with a large proportion of minority students had lower intercepts and a greater (more negative) gender by stakes interaction. I removed the three-way interaction and the non-significant two-way interactions from the model. I also deleted the proportion-minority predictor for all effects except the intercept and gender by stakes interaction. In this model (not shown), the proportion-minority no longer significantly influenced the gender by stakes interaction ($p = .785$), so I removed it from the final model summary in Table 6.

These more parsimonious models are shown in Tables 5 and 6.

Table 5 - Log-Odds of Response on Science CR Items, the Final Model

	Coefficient	Standard Error	Approx. T-ratio	d.f	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	2.781400	0.115269	24.130	26	0.000
MINORITY, γ_{001}	-1.633169	0.609339	-2.680	26	0.013
For STAKES, β_{01k}					
INTERCEPT3, γ_{010}	0.633131	0.146497	4.322	3991	0.000
MINORITY, γ_{011}	0.962471	0.267452	3.599	3991	0.001
For ETHNICITY, β_{02k}					
INTERCEPT3, γ_{020}	-0.503269	0.216514	-2.324	3992	0.020
For GENDER, β_{03k}					
INTERCEPT3, γ_{030}	-0.555828	0.076539	-7.262	3991	0.000
MINORITY, γ_{031}	0.474410	0.134265	3.533	3991	0.001

Table 6 - Log-Odds of Response on Math CR Items, the Final Model

	Coefficient	Standard Error	Approx. T-ratio	d.f	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	2.022953	0.063159	32.029	30	0.000
MINORITY, γ_{001}	-1.624456	0.176036	-9.228	30	0.000
For STAKES (S), β_{01k}					
INTERCEPT3, γ_{010}	1.353810	0.162707	8.321	3468	0.000
For GENDER (G), β_{02k}					
INTERCEPT3, γ_{020}	-0.485008	0.061482	-7.889	3468	0.000
For S X G, β_{03k}					
INTERCEPT3, γ_{030}	-0.335005	0.096537	-3.470	3468	0.001

The coefficients in Tables 5 and 6 are in the log-odds metric. Tables 7 and 8 show the predicted log-odds, odds, and probabilities by gender, stakes, and ethnicity. These would be estimates for a school with an average proportion of non-White students (because school proportion-minority was centered around the mean). When ethnicity was not a factor in the tables, the estimate was for a group of students whose ethnic

background was proportional to the high-stakes sample. The log-odds were found by substituting the coefficients from Tables 5 and 6 into the hierarchical models (using the codings for stakes, gender, and ethnicity detailed earlier), and then the log-odds were mathematically transformed to odds and probabilities. The probability-ratios and odds-ratios shown in the tables are the ratio of the first group to the second group listed for that factor (for example, the gender ratio is the ratio of girls:boys because girls are listed first).

Table 7 - Predicted Log-Odds, Probability (Rate), and Odds of Response, Science

	log-odds	probability	probability ratio	odds	odds ratio
population average	2.78	0.94		16.14	
gender			1.03		1.88
girls	3.05	0.95		21.18	
boys	2.50	0.92		12.15	
stakes			1.02		1.61
high stakes	3.10	0.96		22.15	
low stakes	2.46	0.92		11.76	
ethnicity			0.97		0.60
Black	2.34	0.91		10.37	
White	2.84	0.94		17.16	

Table 8 - Predicted Log-Odds, Probability (Rate), and Odds of Response, Math

	log-odds	probability	probability ratio	odds	odds ratio
population average	2.02	0.88		7.56	
Stakes			1.14		3.87
High Stakes	2.70	0.94		14.88	
Girls	3.02	0.95	1.03	20.48	1.92
Boys	2.37	0.91		10.67	
Low Stakes	1.35	0.79		3.84	
Girls	1.50	0.82	1.05	4.49	1.37
Boys	1.18	0.77		3.27	

In science, girls were 0.03 times more likely (in the probability metric, because 95/92 = 1.03) to respond to each item than boys were (odds ratio of 21.18/12.15 = 1.88). Students taking the test under high stakes conditions were 0.02 times more likely (probability ratio:96/94 = 1.02) to respond than students tested under low stakes conditions (odds ratio of 1.61). Blacks were 0.97 times as likely as Whites to respond to a constructed response item (odds ratio of 0.60).

In math, girls were 0.05 times more likely (main effect not shown in table, probability ratio = 93/88 = 1.05) to respond than boys were (odds ratio of 1.62). Under high stakes, students were 0.14 times more likely to respond than they were under low stakes (odds ratio of 3.87). In the odds metric, the ratio of girls' to boys' response tendencies was greater under high stakes than under low stakes, but in the probability metric this ratio was *slightly* greater under low stakes. This shows the non-linear relationship between the two metrics.

To illustrate how these effects differed with the proportion of minority students, predicted results for a school with 5% minority students are compared with the predicted results for a school with 90% minority students in Tables 9 and 10. Because there were only a small number of schools, the estimates of the effect of the school-level proportion minority are less precise then the estimates of the student-level effects, but they give an approximation of the possible between-school differences.

Table 9 - Predictions for Schools with High and Low Minority Enrollment, Science

5% Minority	log-odds	probability	probability ratio	odds	odds ratio
population average	3.05	0.95		21.07	
gender			1.03		1.88
girls	3.36	0.97		28.71	
boys	2.72	0.94		15.24	
stakes			1.02		1.61
high stakes	3.29	0.96		26.73	
low stakes	2.81	0.94		16.60	
ethnicity			0.97		0.60
Black	2.61	0.93		13.54	
White	3.11	0.96		22.39	

90% Minority	log-odds	probability	probability ratio	odds	odds ratio
population average	1.66	0.84		5.26	
gender			1.04		1.26
girls	1.77	0.85		5.88	
boys	1.54	0.82		4.67	
stakes			1.24		3.65
high stakes	2.31	0.91		10.04	
low stakes	1.01	0.73		2.75	
ethnicity			0.91		0.60
Black	1.22	0.77		3.38	
White	1.72	0.85		5.59	

Table 10 - Predictions for Schools with High and Low Minority Enrollment, Math

5% Minority	log-odds	probability ratio	probability	odds	odds ratio
population average	2.29	0.91		9.85	
Stakes			1.14		3.87
High Stakes	2.96	0.95		19.39	
Girls	3.28	0.96	1.03	26.70	1.92
Boys	2.63	0.93		13.90	
Low Stakes	1.61	0.83		5.01	
Girls	1.77	0.85	1.05	5.85	1.37
Boys	1.45	0.81		4.26	

90% Minority	log-odds	probability ratio	probability	odds	odds ratio
population average	0.92	0.71		2.51	
Stakes			1.48		3.87
High Stakes	1.60	0.83		4.93	
Girls	1.92	0.87	1.12	6.79	1.92
Boys	1.26	0.78		3.54	
Low Stakes	0.24	0.56		1.27	
Girls	0.40	0.60	1.15	1.49	1.37
Boys	0.08	0.52		1.08	

In science, for the effects of test stakes, the odds ratio (high stakes: low stakes) was much larger in the high-minority school (10.04/2.75 = 3.65) than in the low-minority school (26.73/16.60 = 1.61). To a lessor extent, the opposite was true for the gender effect. However, because of the non-linear relationship between the odds and the probability, the probability ratio (girls: boys) was actually slightly larger in the high-minority school though the odds ratio was larger in the low-minority school. [In the second cohort, both the odds ratio and the probability ratio showed greater gender differences in high-minority schools--see Appendix.] While the odds ratio remained unchanged for the ethnic effect (because school proportion-minority did not have a

significant effect on the ethnic effect in the log-odds units), the probability ratio changed when the school-average log-odds changed.

In math, the average log-odds changed significantly with proportion-minority; the log-odds decreased as the proportion-minority increased. The odds-ratios for the other effects did not change significantly with school proportion-minority. However, in the probability metric, stakes made a bigger difference in the high-minority schools. In the low-minority schools, students were 14% more likely to respond under high stakes, while in the high-minority schools, students were 48% more likely to respond under high stakes (while the odds-ratio was 3.87 for both comparisons).

Further attempts were made to add a random school effect to the model for the effects of stakes, gender, and ethnicity (random effects were added to each of these terms individually). In science, the between-school variance in gender effects was not statistically different from zero. With only 28 schools and one random term already in the model (the school mean), the variance could not be estimated for the stakes or ethnicity effects. The problem appeared to be collinearity with the random school intercept. In math, the variance in the school stakes effect was about 44% the size of the variance in school means, and the two effects had a correlation of .36. The variance in gender effects and ethnicity effects could not be estimated while the random school mean effect was in the model.

Ability Scores

Relationship Between Constructed Response and Multiple Choice Scores

All items were calibrated together, using the responses of the high stakes group. This procedure implicitly involves the assumption that both response formats measure one

predominant factor. One way of checking this is Yen's Q3, which was described in the Method section to support the treatment of item clusters as independent items. If the constructed response items measured something consistently different than the multiple choice items, there would be highly correlated residuals among the constructed response items. However, in science all the correlations among the constructed response items were no larger than .04 (slight negative correlations are expected because each item contributes to the ability score), after the two "investigation" items were combined. The constructed response items, then, were not measuring a common factor apart from the predominant factor measured by the test as a whole. In math, on the other hand, 3 of the 15 correlations among the constructed response items were greater than .10, and the other 12 were greater than 0. While none of the correlations were large enough to cause a practical problem for item estimation, these items did seem to measure (to a small degree) something different from the test as a whole, which could indicate either the items were measuring another aspect of math or they were measuring a construct-irrelevant, but shared, source of variance.

A simple way of checking the relationship between the formats is to look at the correlation between the raw scores. In the high stakes group, this correlation was .78 in science and .80 in math. Another way to estimate the correlation is to run a random-effects hierarchical linear model with two outcomes for each student, multiple choice score and constructed response score (these were estimated on the same scale, using a one-parameter item response theory model). With no student-level factors in the model, the correlation between the residuals for these two scores is the estimated within-school correlation between the two "true" scores. The measurement error is taken into

consideration by using the standard error of the ability estimate (from the IRT estimation) in the within-student variance. Using this model, the correlation between the scores was estimated to be about .98 in science, and .92 in math.

The constructed response items did add to the precision of the measurement scale. The information functions for scores based on the multiple choice items alone, compared to scores based on all items, are plotted in Figures 1 and 2.

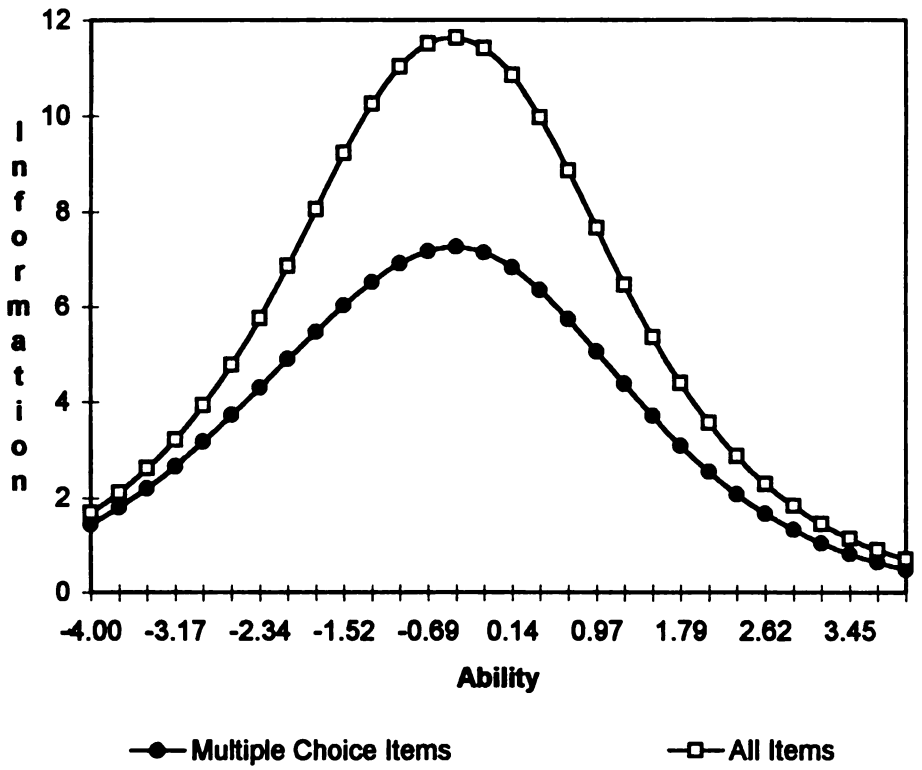

Figure 1: Information Function, Science

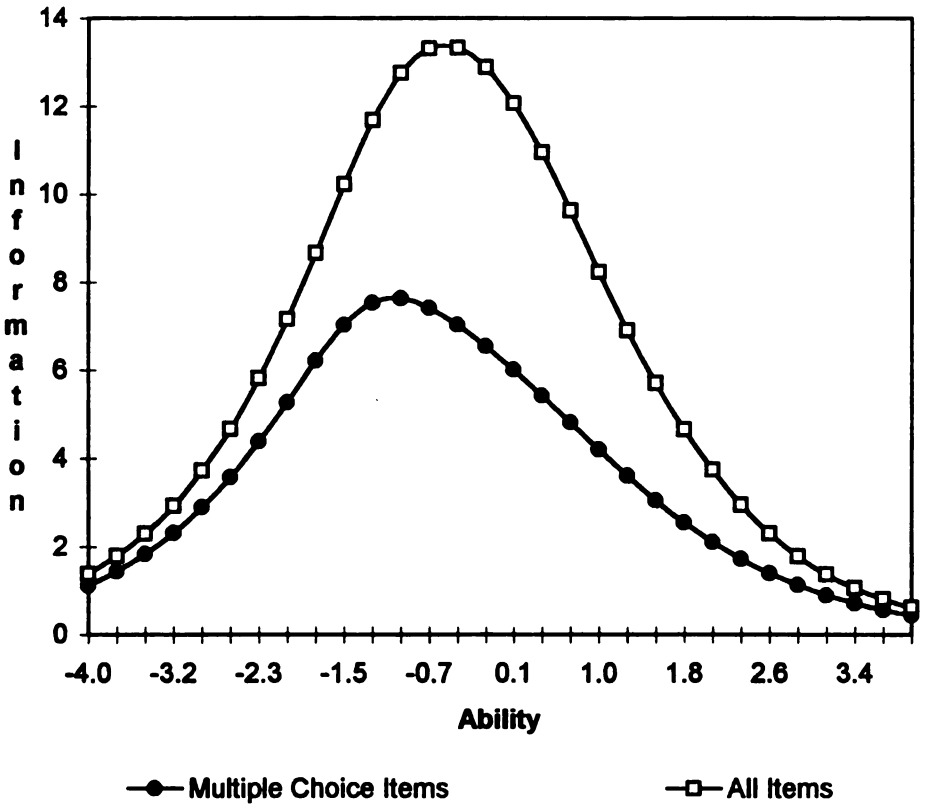

Figure 2: Information Function, Math

Factors Affecting Scores

From these item estimates based on the high-stakes sample, two scores were estimated for each student, one based on the multiple choice items and the other based on the constructed response items. Because both types of items were estimated together, the average score was defined to be 0 on both subscales, in the high-stakes group.

After scores were estimated for all students, the variance due to schools and to students within schools was estimated with a random-effects hierarchical linear model, with the within-student variance set to the square of the estimated standard error of measurement (based on the IRT score estimation--the inverse of the information function at the student's estimated ability). About 25% of the between-student variance in science,

and 23% in math, was due to between-school differences. Next, the proportion of students who identified themselves as non-White was entered into the model to predict each school mean. This characteristic accounted for about 68% of the between-school variance in scores in science, and 48% in math.

Next, the format of the items was added to the model. Format was coded +0.5 for the constructed response score and -0.5 for the multiple choice score, so the intercept would be the mean of the two scores and the coefficient for format would be the difference between the scores. In science, the within-school variance in format effects could not be estimated. At an intermediate estimation step, the estimated variance was not significantly greater than zero (χ^2=3697.57, df=3993). In further models, the format effect was held constant (or constant conditional on student factors) within schools, but allowed to vary across schools. Given that the within-school variance in the format effect seemed to be nearly zero, it might not seem reasonable to attempt to find factors which explained differences in the format effect. However, as there were theoretical reasons to include specific factors (test stakes, gender, and ethnicity) which might be associated with the size of the format effect, these factors were included in later models (and the format effect was held constant within schools, conditional on these factors). [In the second science cohort, there was small but significant variance in format effects, until student-level characteristics were added to the model.]

In science, schools' overall means could be estimated more reliability than their format effects; the reliability was .954 for the means and .552 for the format effects. The reliability was .793 for the student-level means. The correlation between the school mean

and the school format effect was .807. In schools with low means, the constructed response items decreased scores more.

In math, the within-school variance due to format could be estimated, but the reliability of the format effect for individual students was low (.221). The reliability of the school format effect was much higher (.943), as was the reliability for student means (.858) and school means (.917). There was a negative correlation between format effects and means of -.29 at the student level and -.35 at the school level. Those who scored high overall were likely to have a greater decrease due to the constructed response format, the opposite of the finding in science. Across subject areas, the format of the items did not consistently have a greater impact on students of one achievement level.

For the next model, I added the student characteristics. At level-1 (within-students), the format of the items in the subscale was a factor (held constant within schools for science). At level-2, the factors were test stakes, student gender, and ethnicity. They were coded the same as in the models for response rate: high stakes = 0.5, low-stakes = -0.5, male = 0.51, female = -0.49, Black = 0.879 in science and 0.842 in math, White = -0.121 in science and -0.158 in math. These codings allowed the level-3 intercepts to be the average effects for schools with average proportions of genders and ethnicities (because in these schools the average gender and ethnic group would equal zero). The school mean (intercept) and format effect were allowed to vary randomly across schools (proportion of minority was added back later), but the stakes, gender, and ethnicity effects were held constant across schools.

The model was:

$$\hat{\theta}_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{format}) + e_{ijk},$$

where $\hat{\theta}_{ijk}$ is the estimated ability under format i for student j in school k, π_{0jk} is the average of the two formats, π_{1jk} is the difference between the formats, and e_{ijk} is random error (the variance of which was estimated through IRT).

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{stakes}:X_{1jk}) + \beta_{02k}(\text{ethnicity}:X_{2jk}) + \beta_{03k}(\text{gender}:X_{3jk}) + \beta_{04k}(X_{1jk}X_{2jk}) + \beta_{05k}(X_{1jk}X_{3jk}) + \beta_{06k}(X_{2jk}X_{3jk}) + \beta_{07k}(X_{1jk}X_{2jk}X_{3jk}) + r_{0jk},$$

$$\pi_{1jk} = \beta_{10k} + \beta_{11k}(X_{1jk}) + \beta_{12k}(X_{2jk}) + \beta_{13k}(X_{1jk}X_{2jk}) + \beta_{14k}(X_{1jk}X_{2jk}) + \beta_{15k}(X_{1jk}X_{3jk}) + \beta_{16k}(X_{2jk}X_{3jk}) + \beta_{17k}(X_{1jk}X_{2jk}X_{3jk}) + r_{1jk}, \text{ (no } r_{1jk} \text{ for science)}$$

where β_{00k} and β_{10k} are the intercepts for school k's mean and format, effect, respectively, the other β's are the coefficients for the factors (X's) as defined the first time each factor is listed, and the variances of r_{0jk} and r_{1jk} are random within-school variances.

$$\begin{aligned}
\beta_{00k} &= \gamma_{000} + u_{00k}, & \beta_{10k} &= \gamma_{100} + u_{10k}, \\
\beta_{01k} &= \gamma_{010}, & \beta_{11k} &= \gamma_{110}, \\
\beta_{02k} &= \gamma_{020}, & \beta_{12k} &= \gamma_{120}, \\
\beta_{03k} &= \gamma_{030}, & \beta_{13k} &= \gamma_{130}, \\
\beta_{04k} &= \gamma_{040}, & \beta_{14k} &= \gamma_{140}, \\
\beta_{05k} &= \gamma_{050}, & \beta_{15k} &= \gamma_{150}, \\
\beta_{06k} &= \gamma_{060}, & \beta_{16k} &= \gamma_{160}, \\
\beta_{07k} &= \gamma_{070}, & \beta_{17k} &= \gamma_{170},
\end{aligned}$$

where the γ's are intercepts (averages across schools) for their respective β's and the variances of u_{00k} and u_{10k} are random between-school variances.

The estimates for this model are shown in Tables 11 and 12. Again, level-3 coefficients are shown under the level-2 effects they predict, which in turn are listed under the associated level-1 effects. Thus, all predictors of student j's (in school k) intercept (π_{0jk}) are listed before the predictors of the student's format effect (π_{1jk}). The effects listed

under π_{1jk} essentially represent interactions with format because they modify the format effect (except the intercept, γ_{100}, which is the main effect of format).

Table 11 - Student-Level Predictors for Ability Scores, Science

	Coefficient	Standard Error	Approx. T-ratio	d.f.	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	-0.219714	0.046325	-4.743	27	0.000
For STAKES (S), β_{01k}					
INTERCEPT3, γ_{010}	0.266526	0.054286	4.910	3964	0.000
For ETHNICITY (E), β_{02k}					
INTERCEPT3, γ_{020}	-0.441193	0.091912	-4.800	3964	0.000
For GENDER (G), β_{03k}					
INTERCEPT3, γ_{030}	0.050724	0.018751	2.705	3964	0.007
For S X E, β_{04k}					
INTERCEPT3, γ_{040}	-0.037600	0.079854	-0.471	3964	0.637
For S X G, β_{05k}					
INTERCEPT3, γ_{050}	0.046886	0.047226	0.993	3964	0.321
For E X G, β_{06k}					
INTERCEPT3, γ_{060}	-0.034031	0.046008	-0.740	3964	0.459
For S X G X E, β_{07k}					
INTERCEPT3, γ_{070}	-0.129336	0.180275	-0.717	3964	0.473
For FORMAT slope, π_{1jk}					
For INTERCEPT2, β_{10k}					
INTERCEPT3, γ_{100}	-0.097831	0.016483	-5.935	27	0.000
For STAKES, β_{11k}					
INTERCEPT3, γ_{110}	0.187509	0.022731	8.249	4020	0.000
For ETHNICITY, β_{12k}					
INTERCEPT3, γ_{120}	-0.041197	0.037648	-1.094	4020	0.274
For GENDER, β_{13k}					
INTERCEPT3, γ_{130}	-0.182170	0.018896	-9.641	4020	0.000
For S X E, β_{14k}					
INTERCEPT3, γ_{140}	0.038476	0.074645	0.515	4020	0.606
For S X G, β_{15k}					
INTERCEPT3, γ_{150}	0.017272	0.039639	0.436	4020	0.663
For E X G, β_{16k}					
INTERCEPT3, γ_{160}	0.036570	0.083583	0.438	4020	0.661
For S X G X E, β_{17k}					
INTERCEPT3, γ_{170}	-0.036484	0.071973	-0.507	4020	0.612

Table 12 - Student-Level Predictors for Ability Scores, Math

	Coefficient	Standard Error	Approx. T-ratio	d.f.	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	-.292614	0.041978	-6.971	31	0.000
For STAKES (S), β_{01k}					
INTERCEPT3, γ_{010}	0.476044	0.082057	5.801	3436	0.000
For ETHNICITY (E), β_{02k}					
INTERCEPT3, γ_{020}	-0.699293	0.090220	-7.751	3436	0.000
For GENDER (G), β_{03k}					
INTERCEPT3, γ_{030}	-0.031611	0.038610	-0.819	3436	0.413
For S X E, β_{04k}					
INTERCEPT3, γ_{040}	0.079787	0.100327	0.795	3436	0.427
For S X G, β_{05k}					
INTERCEPT3, γ_{050}	0.086615	0.048598	1.782	3436	0.074
For E X G, β_{06k}					
INTERCEPT3, γ_{060}	-0.097234	0.102314	-0.951	3436	0.342
For S X G X E, β_{07k}					
INTERCEPT3, γ_{070}	0.041274	0.120575	0.342	3436	0.732
For FORMAT slope, π_{1jk}					
For INTERCEPT2, β_{10k}					
INTERCEPT3, γ_{100}	-0.239405	0.050330	-5.830	31	0.000
For STAKES, β_{11k}					
INTERCEPT3, γ_{110}	0.477970	0.100358	4.763	3436	0.000
For ETHNICITY, β_{12k}					
INTERCEPT3, γ_{120}	0.130640	0.076187	1.715	3436	0.086
For GENDER, β_{13k}					
INTERCEPT3, γ_{130}	-0.182232	0.026402	-6.902	3436	0.000
For S X E, β_{14k}					
INTERCEPT3, γ_{140}	0.212012	0.134392	1.578	3436	0.114
For S X G, β_{15k}					
INTERCEPT3, γ_{150}	-0.017074	0.052135	-0.328	3436	0.743
For E X G, β_{16k}					
INTERCEPT3, γ_{160}	-0.034828	0.060627	-0.574	3436	0.565
For S X G X E, β_{17k}					
INTERCEPT3, γ_{170}	-0.025654	0.136738	-0.188	3436	0.851

None of the interactions among the student-level characteristics were significant at the .05 level (though in math the stakes by gender effect was close (γ_{050} = .087, p = .074)), but each of the student-level characteristics had a significant main effect (except gender in math, γ_{030}) and stakes and gender influenced (interacted with) the response format effect (γ_{110} and γ_{130}). When students took the test under high stakes, their subscores were defined to be approximately equal on average. Under low-stakes, however, the multiple choice score was higher than the constructed response score. In science, for boys the average constructed response score was lower than the average multiple choice score, while for girls the two scores were, on average, about the same. In math, the average constructed response score was lower for both genders, but the difference was larger (more negative) for boys. Within schools, the average score for Blacks was lower than the average score for Whites.

Next, the interactions among the student-level factors (stakes, ethnicity, and gender) were removed from the model for science. Removing all these effects simultaneously decreased the number of parameters estimated by eight, and increased the deviance of the model by about 4.209. The deviance is a measure of fit (or misfit); it is distributed as chi-square with degrees of freedom equal to the number of parameters, so the change in deviance is approximately distributed as chi-square with degrees of freedom equal to the change in number of parameters (Bryk & Raudenbush, 1992, Chapter 3). The fit of the model, then, did not significantly change with the removal of these eight interaction terms ($\chi^2_{(8)}$ = 4.209, p = .838).

For math, all the interactions among the student level factors, except stakes by gender, were removed. The stakes by gender effect became less significant when these

other effects were removed (and remained so in later attempts at models including the school proportion-minority), so it was removed also. The change in deviance between the models with all student-level interactions and the model with none of these interactions was 15.377, again with 8 degrees of freedom. This was almost significant at the .05 level, so removing the set of interactions as a whole did lead to somewhat worse fit. However, none of the individual interactions could be pinpointed as being particularly useful.

For the next model, the proportion of students who identified themselves as non-White was used as a predictor at the school level. This predictor was centered at the grand mean, so the intercepts would be the values for a school with an average proportion of non-White students (about 21-22%). The estimated coefficients of this model are displayed in Tables 13 and 14.

In science, proportion of minority students was a significant predictor only of the school intercept ($\gamma_{001} = -0.60$, $p = .025$). [In the second cohort, proportion of minority students was a significant predictor only of the gender and gender by format effects]. For the final model (Table 15), then, proportion of minority students was removed as a predictor except for this one effect. Notice this change produced a change in the ethnicity effect on the format effect (the ethnicity by format interaction, γ_{120}). With proportion of minority students as a school level predictor of all effects, the ethnicity by format effect was slightly positive and not significantly different from zero (see Table 13). With proportion of minority students as a predictor only of the mean, the ethnicity by format effect was significantly negative and its standard error decreased by half (see Table 15). When proportion of minority students was removed (except for as a predictor of the

mean), seven fewer parameters were estimated and the deviance of the model changed by only 4.03 (not a statistically significant change).

Table 13 - Predicted Scores in Science, Controlling School Minority Composition

	Coefficient	Standard Error	Approx. T-ratio	d.f.	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	-0.224317	0.051854	-4.326	26	0.000
MINORITY, γ_{001}	-0.604706	0.254938	-2.372	26	0.025
For STAKES, β_{01k}					
INTERCEPT3, γ_{010}	0.267159	0.052223	5.116	3963	0.000
MINORITY, γ_{011}	0.040849	0.101976	0.401	3963	0.688
For ETHNICITY, β_{02k}					
INTERCEPT3, γ_{020}	-0.391861	0.127932	-3.063	3963	0.003
MINORITY, γ_{021}	0.116587	0.278516	0.419	3963	0.675
For GENDER, β_{03k}					
INTERCEPT3, γ_{030}	0.060345	0.016665	3.621	3963	0.001
MINORITY, γ_{031}	-0.028359	0.037500	-0.756	3963	0.450
For FORMAT slope, π_{1jk}					
For INTERCEPT2, β_{10k}					
INTERCEPT3, γ_{100}	-0.090846	0.024023	-3.782	26	0.001
MINORITY, γ_{101}	-0.120986	0.119589	-1.012	26	0.321
For STAKES, β_{11k}					
INTERCEPT3, γ_{110}	0.189326	0.023235	8.148	4019	0.000
MINORITY, γ_{111}	0.051586	0.093655	0.551	4019	0.581
For ETHNICITY, β_{12k}					
INTERCEPT3, γ_{120}	0.019063	0.064370	0.296	4019	0.767
MINORITY, γ_{121}	-0.075475	0.144707	-0.522	4019	0.602
For GENDER, β_{13k}					
INTERCEPT3, γ_{130}	-0.179031	0.020228	-8.851	4019	0.000
MINORITY, γ_{131}	-0.007193	0.123969	-0.058	4019	0.954

Table 14 - Predicted Scores in Math, Controlling School Minority Composition

	Coefficient	Standard Error	Approx. T-ratio	d.f.	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	-0.274292	0.069874	-3.925	30	0.001
MINORITY, γ_{001}	0.005295	0.417567	0.013	30	0.990
For STAKES (S), β_{01k}					
INTERCEPT3, γ_{010}	0.476844	0.081774	5.831	3435	0.000
MINORITY, γ_{011}	0.140810	0.138483	1.017	3435	0.310
For ETHNICITY, β_{02k}					
INTERCEPT3, γ_{020}	-0.643642	0.112995	-5.696	3435	0.000
MINORITY, γ_{021}	-0.188647	0.494516	-0.381	3435	0.702
For GENDER (G), β_{03k}					
INTERCEPT3, γ_{030}	-0.026184	0.038278	-0.684	3435	0.494
MINORITY, γ_{031}	0.003311	0.118061	0.028	3435	0.978
For S x G, β_{04k}					
INTERCEPT3, γ_{040}	0.074895	0.047568	1.574	3435	0.115
MINORITY, γ_{041}	-0.145633	0.161804	-0.900	3435	0.368
For FORMAT slope, π_{1jk}					
For INTERCEPT2, β_{10k}					
INTERCEPT3, γ_{100}	-0.328920	0.051302	-6.411	30	0.000
MINORITY, γ_{101}	-0.331858	0.186887	-1.776	30	0.085
For STAKES, β_{11k}					
INTERCEPT3, γ_{110}	0.477706	0.101277	4.717	3435	0.000
MINORITY, γ_{111}	0.265831	0.149640	1.776	3435	0.075
For ETHNICITY, β_{12k}					
INTERCEPT3, γ_{120}	0.156460	0.099890	1.566	3435	0.117
MINORITY, γ_{121}	0.346326	0.276510	1.252	3435	0.211
For GENDER, β_{13k}					
INTERCEPT3, γ_{130}	-0.184764	0.027467	-6.727	3435	0.000
MINORITY, γ_{131}	-0.052860	0.098589	-0.536	3435	0.591

In math, school proportion-minority was a borderline significant predictor of the format effect (γ_{101}) and the stakes by format interaction (γ_{111}). After removing it as a predictor for all other effects, and removing the stakes by gender effect (as described above), eight fewer parameters were estimated and the deviance changed by 6.363 (non-significant).

Table 15 - The Final Model for Science Ability Scores

	Coefficient	Standard Error	Approx. T-ratio	d.f.	P-value
ForINTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	-0.214849	0.037118	-5.788	26	0.000
MINORITY, γ_{001}	-0.473096	0.149135	-3.172	26	0.004
For STAKES, β_{01k}					
INTERCEPT3, γ_{010}	0.267049	0.053331	5.007	3964	0.000
For ETHNICITY, β_{02k}					
INTERCEPT3, γ_{020}	-0.380583	0.100030	-3.805	3964	0.000
For GENDER, β_{03k}					
INTERCEPT3, γ_{030}	0.060361	0.017315	3.486	3964	0.001
For FORMAT slope, π_{1jk}					
For INTERCEPT2, β_{10k}					
INTERCEPT3, γ_{100}	-0.095642	0.015553	-6.149	27	0.000
For STAKES, β_{11k}					
INTERCEPT3, γ_{110}	0.188382	0.023123	8.147	4020	0.000
For ETHNICITY, β_{12k}					
INTERCEPT3, γ_{120}	-0.081150	0.032538	-2.494	4020	0.013
For GENDER, β_{13k}					
INTERCEPT3, γ_{130}	-0.177916	0.019134	-9.299	4020	0.000

Table 16 - The Final Model for Math Ability Scores

	Coefficient	Standard Error	Approx. T-ratio	d.f.	P-value
ForINTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	-0.289806	0.042248	-6.860	31	0.000
For STAKES, β_{01k}					
INTERCEPT3, γ_{010}	0.473559	0.084870	5.580	3436	0.000
For ETHNICITY, β_{02k}					
INTERCEPT3, γ_{020}	-0.673611	0.089103	-7.560	3436	0.000
For GENDER, β_{03k}					
INTERCEPT3, γ_{030}	-0.016969	0.037698	-0.450	3436	0.652
For FORMAT slope, π_{1jk}					
For INTERCEPT2, β_{10k}					
INTERCEPT3, γ_{100}	-0.293539	0.050224	-5.845	30	0.000
MINORITY, γ_{101}	-0.149719	0.073394	-2.040	30	0.050
For STAKES, β_{11k}					
INTERCEPT3, γ_{110}	0.477662	0.101225	4.719	3435	0.000
MINORITY, γ_{111}	0.281601	0.145959	1.929	3435	0.053
For ETHNICITY, β_{12k}					
INTERCEPT3, γ_{120}	0.171526	0.088514	1.938	3436	0.052
For GENDER, β_{13k}					
INTERCEPT3, γ_{130}	-0.183307	0.027994	-6.548	3436	0.000

The final models are shown in Tables 15 and 16. In both math and science, students scored higher under high stakes (γ_{010}--the estimated difference was greater for math). White students scored higher than Black students (γ_{020}), but in science some of this difference was due to between-school differences in school proportion-minority (γ_{001}). There was a small but significant gender difference in science (γ_{020}), favoring males. In both subject areas gender moderated the response format effect (γ_{130}), with males scoring higher on multiple choice and females scoring relatively higher on constructed response. Within schools, the ethnicity by format interaction (γ_{120}) was in opposite directions in math and the first science cohort, and in math, the school percent minority (γ_{001}) had a different

effect on the format effect than the individual student's ethnicity did (γ_{120}). Students from high minority schools had a larger (more negative) format effect, but within schools the average format effect was larger for Whites. [The effect in the second science cohort was similar to that in math].

Using these coefficients (and the codings for format, stakes, ethnicity, and gender described above), mean scores can be estimated *within a typical school* (one with about 21-22% minority students). These mean scores are shown in Tables 17 and 19. Tables 18 and 20 provide standardized differences for the significant effects. For these tables, the differences were divided by the standard deviation in the entire sample (0.745 in science and 0.891 in math). This allowed for comparison of effects within this study; if the gender groups differed by the same non-standardized value as the ethnic groups differed by, for example, the standardized difference would be the same for both gender and ethnicity. When comparing results from this study with other research, readers might wish to calculate an effect size based on the pooled within-group standard deviation. The standard deviations provided in Tables 17 and 19, with the group sizes reported in Chapter 3, can be used for this purpose. The standard deviations in Tables 17 and 19 were calculated from the variance of the entire sample; they are not within-school variances. Also, they were based on the reliability-weighted scores (scores weighted by the inverse of the square of the standard error of measurement for that score), as the group means were.

Table 17 - Average Predicted Scores in Science

	multiple choice		constructed response		mean	
	mean	sd	mean	sd	mean	sd
average	-0.17	0.72	-0.26	0.78	-0.21	0.68
high stakes	-0.08	0.71	-0.08	0.77	-0.08	0.67
low stakes	-0.25	0.73	-0.44	0.77	-0.35	0.68
Blacks	-0.47	0.59	-0.63	0.74	-0.55	0.58
Whites	-0.13	0.69	-0.21	0.74	-0.17	0.65
Boys	-0.09	0.75	-0.28	0.80	-0.18	0.71
Girls	-0.24	0.68	-0.25	0.77	-0.24	0.66

Table 18 - Standardized Within-School Effects in Science

	standardized difference		
	multiple choice	constructed response	average
stakes	0.23	0.48	0.36
ethnicity	-0.46	-0.57	-0.51
gender	0.20	-0.04	0.08

Table 19 - Average Predicted Scores in Math

	multiple choice		constructed response		mean	
	mean	sd	mean	sd	mean	sd
average	-0.14	0.94	-0.44	0.83	-0.29	0.78
high stakes	-0.03	0.97	-0.08	0.76	-0.05	0.77
low stakes	-0.26	0.88	-0.79	0.80	-0.53	0.75
Blacks	-0.78	0.66	-0.93	0.85	-0.86	0.66
Whites	-0.02	0.94	-0.34	0.80	-0.18	0.76
Boys	-0.10	0.99	-0.49	0.85	-0.30	0.81
Girls	-0.18	0.90	-0.38	0.80	-0.28	0.76

	5% minority			90% minority		
	multiple choice	constructed response	mean	multiple choice	constructed response	mean
average	-0.16	-0.42	-0.29	-0.09	-0.49	-0.29
high stakes	-0.03	-0.08	-0.05	-0.02	-0.08	-0.05
low stakes	-0.29	-0.77	-0.53	-0.16	-0.89	-0.53
Blacks	-0.80	-0.92	-0.86	-0.73	-0.98	-0.86
Whites	-0.04	-0.33	-0.18	0.03	-0.39	-0.18
Boys	-0.12	-0.48	-0.30	-0.05	-0.54	-0.30
Girls	-0.19	-0.37	-0.28	-0.13	-0.43	-0.28

Table 20 - Standardized Within-School Effects in Math

	standardized difference		
	multiple choice	constructed response	average
stakes	0.26	0.80	0.53
ethnicity	-0.85	-0.66	-0.76
gender	0.08	-0.12	-0.02

In a school which had a 5% minority composition, the mean in science would be predicted to increase by 0.054. In a school with a minority enrollment of 90%, the mean would be predicted to decrease by 0.325. In math, predicted scores are shown separately

for a school with 5% minority enrollment and a school with 90% minority enrollment, because this factor was associated with (interacted with) changes in other factors. Note that means are adjusted for student-level ethnicity and gender (except when each is the target variable), so the mean score for the 90% minority school is an estimate of the mean if the school effect related to proportion-minority stayed the same while the ratio of Black to White students was adjusted to reflect the proportions in the total sample. This makes more sense if one regards the school proportion-minority as merely a proxy variable for school and community resources.

The lower means of science scores in high-minority schools applied to both Black and White students within the schools, but the mean of the Black students in the total sample was more influenced by the low school means in the high-minority schools. The within-school ethnic difference was smaller than the difference between the total sample means of the two ethnic groups. Using the reliability-weighted means for the groups in the total sample, the standardized ethnic difference was -0.87 on the multiple choice score, -0.95 on the constructed response score, and -0.89 on the average score. To a lesser degree, because the school proportion minority influenced the format effect in math, the format by ethnicity interaction in math was slightly smaller in the total sample than within-schools. The standardized ethnic difference was -0.83 on the multiple choice score, -0.67 on the constructed response score, and -0.75 on the average score. Again, the standardized differences were computed with the total-sample standard deviation, not the pooled within-group standard deviation (the standard deviations in Tables 17 and 19 can be used to calculate pooled within-group standard deviations).

Next, each school was allowed to have a random effect on the test stakes effect. In science, the variance of these random terms was slightly higher than the variance of the school means, and both were almost four times as high as the variance in school format effects (see Table 21). When random school effects for the ethnicity and gender effects were added to the science model (individually), the estimation algorithm failed to converge in 100 iterations. At that point, estimates suggested that the random effects variance was not significantly greater than zero. In math, the variance in the school stakes effects was higher than the variance in school means but lower than the variance in school format effects (Table 22). The variance in school gender effects was smaller than the other variances. A model with a random term for the school ethnicity effect (including random means and format effects) failed to converge, and intermediate steps indicated the variance was not significantly greater than zero.

Table 21 - Variance Components for Science Ability

Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTERCEPT1/INTERCEPT2, U0	0.18020	0.03247	26	258.5801	0.000
INTERCEPT1/ STAKES, U1	0.19291	0.03722	27	100.2475	0.000
FORMAT/INTERCEPT2, U2	0.04986	0.00249	27	54.5586	0.002

Table 22 - Variance Components for Math Ability

Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTERCEPT1/INTERCEPT2, U0	0.22282	0.04965	30	333.6813	0.000
INTERCEPT1/ GENDER, U1	0.29751	0.08851	30	166.6610	0.000
FORMAT/INTERCEPT2, U2	0.32318	0.10445	30	1505.8560	0.000
INTERCEPT1/INTERCEPT2, U0	0.22059	0.04866	30	392.8767	0.000
INTERCEPT1/ STAKES, U1	0.13379	0.01790	30	59.2044	0.001
FORMAT/INTERCEPT2, U2	0.32407	0.10502	30	1510.5385	0.000

Fit of the High-Stakes Responses to Low-Stakes Item Estimates

Person-Fit

Item parameters were re-estimated from the low-stakes-students' responses. All items were calibrated together, and separate estimations were made for boys and girls. Estimates were made twice, once treating omitted items as incorrect and again treating omitted items as not-administered. Each high-stakes-student's ability was re-estimated based on these item estimates (using only the gender appropriate item estimates, and always scoring omitted items incorrect regardless of how omits were scored for the item estimates). The high stakes-students' fit to these item estimates was measured by the standardized appropriateness index of Levine, Drasgow, and Williams (1985), a composite index based on how probable the student's response to each polychotomous item is, given the student's estimated ability and the item's estimated parameters. This index has been shown to be roughly normally distributed in actual use (Levine, Drasgow, & Williams,

1985). Values less than zero indicate lower than expected fit, and values less than -2.58 would be rare, occurring for only 0.5% of the students in a sample where the model fit. Table 23 shows the mean standardized appropriateness fit, as well as the percent of students with indices less than -2.58, for each condition.

Table 23 - Appropriateness Fit Index

	Girls		Boys	
Science	mean	% <-2.58	mean	% <-2.58
omitted items scored as incorrect	-0.28	1.77%	-0.11	0.43%
omitted items treated as not-administered	-0.25	1.43%	-0.05	0.36%
Math				
omitted items scored as incorrect	-0.43	8.06%	-0.53	9.07%
omitted items treated as not-administered	-0.24	7.09%	-0.25	7.26%

In science, the low-stakes item estimates fit better for the boys than for the girls. Fewer students were judged as misfitting when the item estimates were obtained by treating the omits as not-administered in the low-stakes test (though they were treated as incorrect on the high-stakes test). In math, the low-stakes item estimates fit better for girls than for boys, especially when omitted items were scored as incorrect. For both genders, item estimates fit better when omits were treated as not-administered. However, all sets had very large numbers of misfitting persons.

The appropriateness index was also calculated using the operational item estimates (ignoring gender), to serve as a baseline. The distributions were not precisely normal: In science, the standard deviation was 0.90, the skew was -0.50, and the kurtosis was 0.48.

In math, the standard deviation was 0.99 (essentially 1), the skew was -0.61, and the kurtosis was 0.46. The mean was 0.01 in science and 0.04 in math, and only 0.7% of the students in science and 1.60% in math had values less than -2.58 (somewhat more than the 0.5% expected if normally distributed).

Item-Fit

The appropriateness index targets students whose responses poorly fit the expected pattern. Another index, OUTFIT (Wright & Masters, 1982), targets items which fit the model poorly. The expected value of this index is one. Values greater than one indicate many responses are unexpected, and values less than one indicate responses are "too predictable"--there is not as much unexplained variance as expected.

The four sets of item estimates (gender by method of treating omits) based on the pilot data and responses from the operational data were used to generate four OUTFIT indices for each item. For every set, one science item (a constructed response item) had OUTFIT greater than 1.5. OUTFIT for this item was larger for girls than boys, and it was larger when omits were treated as missing (not wrong) in the pilot data (recall that omits were always treated as wrong in the operational data). In math, two constructed response items had OUTFIT greater than 1.5 for all four sets, a third constructed response item had OUTFIT of 1.46 for one group and above 1.5 for the other three, and the multiple choice item composed of four individual items had OUTFITs between 1.28 and 1.52. Averages for each of the four sets, by item type, are displayed in Table 24.

Table 24 - Average OUTFIT

	Boys				Girls			
	omits as incorrect		omits as missing		omits as incorrect		omits as missing	
Science	mean	sd	mean	sd	mean	sd	mean	sd
Multiple Choice	0.9827	0.1894	0.9884	0.1972	1.0296	0.1306	1.0371	0.1391
Constructed Response	0.9998	0.2971	1.0631	0.2375	1.0407	0.2922	1.0984	0.2457
Math								
Multiple Choice	1.0445	0.1637	1.0051	0.1400	1.0646	0.1566	1.0387	0.1296
Constructed Response	1.6634	0.8989	1.4258	0.5221	1.3589	0.5250	1.2876	0.4455

A repeated-measures ANOVA, with item as the unit of analysis, was conducted on the OUTFIT measures. There were two repeated measures, gender and method of treating omits, and one between factor, item type.

Table 25 - ANOVA Summary Table for OUTFIT, Science

Source	SS	MS	df	F	prob. F
Item Type (IT)	0.058	0.058	1	0.46	0.4999
error (between subjects)	4.592	0.124	37		
Not administered/Incorrect (NA)	0.018	0.018	1	17.03	0.0002
NA X IT	0.025	0.025	1	24.00	<0.0001
error	0.038	0.001	37		
Gender (G)	0.044	0.044	1	2.96	0.0935
G X IT	0.008	0.008	1	0.05	0.8194
error	0.555	0.015	37		
NA X G	0.000	0.000	1	0.08	0.7772
NA X G X IT	0.000	0.000	1	1.32	0.2573
error	0.003	0.000			

Table 26 - ANOVA Summary Table for OUTFIT, Math

Source	SS	MS	df	F	prob. F
Item Type (IT)	3.114	3.114	1	12.12	0.0014
error (between subjects)	8.475	0.257	33		
Not administered/Incorrect (NA)	0.174	0.174	1	9.38	0.0043
NA X IT	0.074	0.074	1	3.97	0.0546
error	0.612	0.019	33		
Gender (G)	0.188	0.188	1	6.09	0.0190
G X IT	0.306	0.306	1	9.92	0.0035
error	1.019	0.031	33		
NA X G	0.040	0.040	1	10.70	0.0025
NA X G X IT	0.029	0.029	1	7.73	0.0089
error	0.124	0.004	33		

In science, the method of treating omits effect and the method by item format interaction were statistically significant. Misfit was higher when omits were treated as not-administered (opposite the finding for the appropriateness index), and the gap was larger on constructed response items than on multiple choice items. Gender did not have a significant main effect on fit, nor did it interact with method of treating omits or item type. The trend was for OUTFIT to be greater for girls, but this effect was not statistically significant.

In math, the three-way interaction among way of treating omits, gender, and item type was statistically significant. For both genders, there was an interaction between treatment of omits and response format; treating omits as not-administered improved fit more for the constructed response items than for the multiple choice items. This interaction was more extreme for boys than for girls.

Both the person-fit and the item-fit were extremely poor in math. One possible reason is because the equal-slopes model was more problematic in math than it was in science. Half-points were possible for the constructed-response items, and in order to

utilize all information from these half-points and yet avoid overweighting the constructed-response items, the slope of each constructed-response item was set to half the slope of each multiple choice item for this study (the slope determines the weight of the item, and each category of the constructed response items was to be weighted a half point). To check the appropriateness of this ratio, a two-parameter model was run, using the high-stakes data. When the slopes were free to vary, the ratio of the average constructed response slope to the average multiple choice slope was about 0.6 in science, fairly close to the imposed 0.5 ratio. In math, however, this ratio was approximately 0.3. This probably contributed to the poor fit. The average OUTFIT for the constructed response items based on the operational item estimates was 1.265, which was better than the OUTFIT based on the pilot item estimates but still quite high for an average. However, the person-fit of the operational data using operational item estimates, as described above, was considerably better. Much of the poor fit, therefore, seems to be due to differences between pilot and operational responses.

CHAPTER 5

DISCUSSION/SUMMARY, IMPLICATIONS, AND CONCLUSIONS

Discussion/Summary

Response Rate: Science

On the constructed response items, the stakes of the test, student ethnicity, and student gender were associated with response rate, with the school mean response rate controlled. Students were more likely to respond under high stakes, girls were more likely to respond than boys, and Whites were more likely to respond than Blacks. These factors did not interact with each other. The higher response rate under high stakes suggests that some of the non-response under low-stakes was due to lack of motivation and not solely to lack of knowledge. The lack of interaction between stakes and ethnicity or gender suggests that, within a school, Blacks and Whites, and boys and girls, are equally affected by test stakes and the ethnic and gender differences do not change with the test stakes.

The ethnic composition of the school was significantly associated with the effects of test stakes and gender, but not with the effects of ethnicity. Schools with more minority students had a larger stakes effect. In schools with high minority enrollments, students appear to be particularly unmotivated (as evidenced by response rate) on low-stakes tests. When the minority enrollment was *not* included as a factor, and the stakes effect was held constant at the average across schools, there was an interaction between ethnic group and test stakes. This was apparently due to between-school differences, not

to ethnic differences within schools, because it disappeared when school minority enrollment was added to the model. [Though in the second cohort there appeared to be some stakes by ethnicity interaction within schools as well]. Looking at the means for each condition, the response rate of Blacks on the high stakes test would appear to be underestimated (relative to the scores of Whites) by the low stakes test when all schools are pooled together.

Response Rate: Math

In math, students were more likely to leave constructed response items blank than they were in science. This suggests students might have had more difficulty thinking of a relevant solution to the math constructed response items, and were not just poorly motivated. Or perhaps many students are more familiar with making written verbal responses in science than in math. In math, their experiences with constructed response items might be limited to showing their computations, while the HSPT requires students to explain their reasoning and solutions.

The stakes of the test, gender, and the school's ethnic composition were all related to tendency to respond to the constructed response items. Students were more likely to respond under high stakes, and girls were more likely to respond than boys. Students in schools with many non-White students were less likely to respond than students in mostly-White schools. Once this effect was controlled, ethnicity was not associated with response rate; the sizable differences between the means for Blacks and Whites could be attributed to between-school differences. This is in contrast to the findings in science, where there were still some within-school ethnic group differences. However, any conclusions about ethnic effects at the student level should be tentative because of the

large number of students who did not provide their ethnicity on the high stakes test and thus were not included, especially in the schools selected for the math analysis.

The results in math also differed from the results in science in that the school proportion of minority students did not interact with stakes or gender when the outcome was in the log-odds metric (there appeared to be a stakes by proportion minority interaction using the probability metric, but, as noted, this was not tested for statistical significance). Schools with high non-White enrollment did not consistently differ from mostly-White schools in the degree of the stakes or gender effect (when using the log-odds metric), only in the school mean effect (intercept).

Ability Scores: Science

The stakes of the test, student ethnicity, and student gender were also associated with test scores. Students scored higher under high stakes, Whites scored higher than Blacks, and boys scored slightly higher than girls. These effects, though, all significantly interacted with response format. Because the residual variance of the format effect within schools seemed to be very small (not significantly different from zero), these student-level characteristics would not be expected to influence the format effect. If the formats were measuring the same thing, as suggested by the lack of variance in format differences, there would be no reason to think gender, ethnicity, or stakes would interact with format. The evidence that the format effect does depend on stakes, gender, and ethnicity suggests that the two formats are not measuring exactly the same thing.

High stakes tended to increase scores for both response formats, but the effect was larger for constructed response items. This was expected based on findings from low-stakes tests that motivation is lower on constructed response items (Sundre, 1996;

Wainer, 1993). If motivation was more equal across item format on the high stakes tests, the format effect would be expected to be smaller on the high stakes test, as was found in this study. This is also consistent with Wolf, Smith, and Birnbaum's (1995) finding that "mentally taxing" items were less difficult (relatively) under high stakes.

Whites scored higher on both formats, but the difference was slightly larger on the constructed response scale in the first cohort and slightly smaller on the constructed response scale in the second cohort.

The gender gap changed direction depending on the response format; boys did better on the multiple choice subscale and girls did slightly better (about the same) on the constructed response subscale. Using only the pilot data, and ignoring the school effect, DeMars (in press) found a trend in this direction (significant on one test form but significant only for certain ability groups on another), which is apparently confirmed with the operational data. This interaction has been consistently found on Advanced Placement exams in the sciences (Bridgeman, 1989; Bridgeman & Lewis, 1994; Schmitt, Mazzeo, & Bleistein, 1991).

Schools with high minority enrollment had lower mean scores (in the second cohort, there was a larger ethnic difference within-schools, but not a significant difference between schools when student characteristics were controlled). However, minority enrollment did not significantly influence the other effects (except gender and gender by format in the second cohort). Because the ethnic composition of the school had an effect on differences in response rate due to test stakes, it might also be expected to influence differences in test scores due to test stakes, at least for the constructed response items. While schools with high minority enrollments had a larger increase in response rate under

high stakes than schools with lower minority enrollments, the increase in test scores did not depend on minority enrollment. Under high stakes, test scores increased about equally for schools with many or few minority students.

Ability Scores: Math

The stakes of the test and student ethnicity were associated with test scores. On average, students did better under high stakes, and Whites did better than Blacks. On the total test score, there was essentially no gender difference, but there was an interaction between gender and response format. Girls scored higher on constructed response items and boys scored higher on multiple choice items. This same interaction was found in science. It is interesting that math (calculus) is the one area on the AP exams where this type of interaction is not found (Schmitt, Mazzeo, & Bleistein, 1991). Also, no interaction was found between response format and gender on several math tests of the General Certificate Exam (GCE) in England (Murphy, 1982). However, the AP calculus exams generally do not require much verbal reasoning/explanation on the constructed response items, in contrast to the HSPT. The interaction found here on the HSPT was similar to that found with a sample of Irish students (Bolger & Kellaghan, 1990).

Stakes and ethnicity also interacted with response format (though ethnicity was not quite statistically significant at the .05 level, and many students provided no ethnic information, so this effect should be interpreted cautiously until it is replicated with other samples). As in science, under low-stakes students did more poorly on the constructed response items than the multiple choice items (in the high stakes sample, the differences were set to approximately zero on the ability scale). The ethnic gap within schools slightly decreased on the constructed response items, compared to a slight increase in science (first

cohort). A decrease in ethnic group differences would be consistent with Badger's (1995) findings in math, but in that study the interaction was larger. On the HSPT, I would conclude that ethnic group differences remained largely stable across response format. Also, the effect of ethnicity at the individual level should be interpreted in the context of the effect of school minority composition in math. This effect, while small, was in the opposite direction; as the proportion of minority students increased, the format effect increased (became more negative).

There was also a significant interaction between school minority composition, test stakes, and response format. In schools with many non-White students, students taking the constructed response items under low-stakes scored particularly low. In these schools, constructed response performance increased more under high-stakes conditions than would be expected in an average school. This effect was not found in science. In science, the only effect of school ethnic composition was on mean scores or gender effects, depending on the cohort. The ethnic group differences in math ability scores seemed to be mainly within schools, not between schools (though, again, lack of ethnic information at the student level may have led to bias in estimates of ethnic effects if the remaining students were not representative of their groups). This is particularly interesting given that the ethnic group differences in response tendency in math seemed to be between schools, not within schools.

Comment on the Ethnicity Variables

The largest discrepancies between math and science, and between the two science forms, involved the effects of student ethnicity and school proportion-minority. Some of these differences may have been due to the omission of students who did not identify their

ethnic groups; the type of student not indicating ethnicity may have varied in the different samples. For example, the test score differences between students who did and did not indicate their ethnicity were slightly greater in math than in science. At the school level, ethnic composition was most likely fairly accurate, as it was based on the pilot data (where ethnic identification was much higher) for schools which had low response to the ethnicity item on the operational test. However, only a relatively small number of schools (about 30) were used for each test form, which would tend to lead to less stable results than a larger number of schools. Furthermore, student-level ethnicity and school ethnic composition are correlated, so it is difficult to separate the effects of each.

Pilot Item Fit to Operational Responses

Because boys were less likely to respond to constructed response items, I expected that, for boys, ignoring omitted items on the pilot test when estimating item parameters would lead to item estimates which fit the operational data better than parameters estimated when omitted items were scored zero on the pilot test. In science, ignoring omitted items (treating them as if they had not been administered), improved person-fit (the appropriateness fit index) only slightly, and it led to somewhat poorer item-fit for the constructed-response items. These differences were small, and probably have little practical meaning. Fit (both item-fit and person-fit) was worse for girls than boys, which was an unexpected finding. Because boys seemed to be less motivated on the pilot test (as evidenced by lower response rates on the constructed response items), I had expected them to have more idiosyncratic responses which would lead to poorer item estimates.

In math, my expectations held. Item fit and person fit both were better when omitted items were treated as not-administered, especially for boys and constructed

response items. In math, the fit was generally much worse than it was in science, so there was more opportunity for differences. Also, in math non-response was higher, so method of treating omits would be expected to make more of a difference. The fit of the persons and of the constructed response items, though, regardless of how omits were treated, was remarkably poor. About 13 to 17 times more people than expected had very poor fit (z-scores less than -2.58), and this was mostly due to the poor fit of the constructed response items. The average outfit of the six constructed response items was 1.29 - 1.66 (depending on the group and conditions). For an individual item this would be 30% higher than expected if there were good fit, and for an average across multiple items it is quite high. This poor fit was partly due to the constraint of equal slopes--when the slopes were free to vary, the slopes of the constructed response items were not as steep as those of the multiple choice items.

Limitations

One limitation to this study is that there was no measure of motivation other than tendency to omit constructed response items. As noted earlier, students may omit items for other reasons, such as little or no knowledge of the correct answer. It would be useful to have data on some other measure of motivation to separate these reasons for omission. Also, students likely varied in how important they felt the diploma endorsements were, depending on their interpretations of the messages they received from the schools, parents, other students, and the media about the diploma endorsements. These student beliefs, in turn, had an impact on how motivated the students were.

Another issue was the problem of disentangling student-level and school-level factors. Within schools, ethnic differences in response rate in science were relatively small

(though significant), comparable in size to gender differences, and in math the within-school ethnic differences were not significantly different from zero. However, the differences between high-minority schools and low-minority schools were larger, and a large percentage of the Black students attended high-minority schools. The school is essentially a proxy variable for a composite of school, neighborhood, and family effects. Students can not be randomly assigned to schools, so there is no easy way of assessing what the within-school ethnic effect would be if non-school background variables were controlled.

Determining the proportion of minority students in a school was in itself a problem. On the operational test, 26% of the tested students in the first cohort and 30% of the students in the second cohort did not identify their ethnic group (even more in the selected schools). In some schools, nearly all students responded, while in others none of the students responded. In the pilot test, at least 80% of the students in each school supplied their ethnicity, so these data could be used for schools which had low-response on the operational test. However, this solution was available only because this study design included only the schools which participated in the pilot test. This option would not be available in most situations. Also, this does nothing to help the problem of missing data at the student level. Though score differences between students who answered the ethnicity item and students who did not were not that great, this information was probably not randomly missing. Teachers' tendencies to include this item as they led students through the process of filling out background information could be associated with school factors, and students' tendencies to complete this item when their classmates did not could be associated with other student factors.

The generalization of these findings should be limited to relatively structured tests. In particular, the findings concerning the constructed response items should not be extended to all forms of performance assessments without further research. The constructed response items were intended to elicit scientific/mathematical reasoning and communication skills. However, all situations and materials were presented on paper only, and all responses were in written form. Somewhat different skills and processes may be involved when students interact with real materials, record their own observations, work with others, and communicate verbally.

The content level of the tests also limits the generalizability of the findings. The level is above the "basic-skills" level, but it includes only concepts to which all students should have been exposed by the end of the 10th grade. Results might be different for advanced content or high-achieving students (such as students taking Advanced Placement exams) and different again for low-level, minimal competency exams. This research focused on a cross-section of students and "typical" required high school curricula.

Implications

Implications for Test-Development

The implications for test-development follow directly from the results. When the dependent variable was the test score (rather than response rate), gender did not interact with test stakes. Ethnicity did not interact with test stakes in science, and the interaction in math (school proportion-minority, format, and stakes interacted) was quite small. This suggests that gender and ethnic differences in test scores can be accurately estimated on low stakes tests; all groups increased similar amounts when the stakes of the test increased. If test-developers were trying to predict how a test cut-score would affect

these groups (Blacks and Whites, boys and girls), they would only need to have some idea of how the overall mean would shift. [Of course, if the cut-score were far from the mean this would not be accurate because it is very possible that group differences near the mean are not the same near the tails of the distribution.] The response rates of students in high-minority schools increased more as the test stakes increased in science, but this did not lead to a disproportionate increase in test scores.

All these effects were moderated by the format of the items. If scores were estimated separately for each response type, increasing the stakes of the test would have a bigger effect on the constructed response scores than on the multiple choice scores. The estimated difficulties of constructed response items relative to multiple choice items on low stakes tests, then, will be off somewhat compared to the relative difficulties estimated under high stakes. Therefore, the item difficulties from pilot forms could lead to inaccuracies if used to equate test forms to be administered under higher stakes. Ideally, information from the high stakes administration should be used in the final equating (as is done for the HSPT).

Ethnic differences seemed about the same on both formats, when school ethnic composition and student-level ethnicity were considered simultaneously; in math, the ethnic difference was somewhat larger on the multiple choice items, but in science the difference was larger on the constructed response items. Gender differences varied with the response format. Using only constructed response items, gender differences would appear to slightly favor girls instead. Adding constructed response items of this type would boost the relative position of females.

Broader Implications for Educators and Policy-Makers

In addition to the implications summarized above, there are some broader implications which are supported by the results, yet move a step further. These are the issues which warrant attention on a policy level. One finding which should be of interest to educators and policy-makers is that some of the ethnic differences could be attributed to school-level differences rather than differences between Blacks and Whites in the same schools. If some of the related causal factors in the school or community can be identified, efforts can be directed toward changing these factors on a schoolwide basis.

The use of multiple response formats is another topic highlighted by these findings. One important issue is the question of what each format measures. The high correlation between the formats suggests they measure something similar, and the fact that the items are estimated on one scale and a single score is typically estimated (the subscales were created only for this study) suggests that the educators involved in the test development believe a single predominant factor is being measured. The constructed response items, though, are not completely redundant and do add information in estimating this single score. Both the multiple choice and constructed response items are designed to measure concepts in the *Michigan Essential Goals and Objectives*, and each section contributes something slightly different towards that purpose. If both formats measure slightly different, but relevant and intended constructs, then both are useful. If the construct-irrelevant sources of variance measured by the formats essentially "balance-out", then that would also be beneficial.

Some educators have advocated using performance-type assessments to drive instruction (Pomplun, 1997), though this departs from the traditional purpose of tests.

Though both response formats may measure very similar constructs, teachers may perceive constructed response items as measuring more reasoning and problem-solving, and adjust their instruction and assignments accordingly (Frederiksen, 1984). Frederiksen also suggested that parents, who have influence on what is taught in schools, are much more likely to be aware of the need for teaching the types of problem-solving used in responding to open-ended items if those items are part of the testing system.

Bennett, Rock, and Wang (1991), in the context of the AP computer science exam, also raised the issue of instructional focus in response to testing. While they found the multiple choice and constructed response portions of the test to be highly correlated, they suggested that one justification for keeping the constructed response items was to keep the emphasis in the class on computer programming (the constructed response items assessed students' abilities to write programs).

Stecher and Hamilton (1994) examined changes in instruction in Vermont after the state mandated portfolio assessments for fourth and eighth grade mathematics. Teachers reported they had students spend more time on problem-solving, class discussions, and explaining and writing about their solutions. Teachers also assigned more problems with ill-defined outcomes, used more hands-on materials, and had the students work together more often. Changes were larger for fourth grade teachers than for eighth grade teachers.

One important difference between portfolio assessment and the constructed response items on the HSPT is that, for portfolio assessment, classroom activities are not necessarily preparation for the assessment; they may actually be part of the assessment. At the extreme, if a teacher previously had not assigned any activities which would be appropriate for the portfolios, he or she would have to add at least enough activities to

build a minimal portfolio. In the case of the HSPT, such an extreme teacher would not have to add any activities beyond the two-hour test itself.

The Kansas Mathematics Assessment is more comparable in format to the HSPT. In Kansas, teachers reported increases in emphasis similar to the Vermont teachers (for example, problem-solving, communication, reasoning, estimation, multiple solutions) in response either to the testing program or to the new state curriculum which went with it (Pomplun, 1997). Change was greatest among elementary school teachers and lowest among high school teachers. Teacher attitude toward the test, and district and building-level responses influenced the amount of change.

Cost is also an important issue for statewide testing. Wainer and Lukhele (1997) estimated that a 30-minute essay (in student time) costs about $7 to score. Stecher (1995), using costs from standardized tests, estimated that constructed response tests (with no hands-on activities) cost about 15 times as much as multiple choice tests of comparable length (in testing time). Further, if hands-on activities and trained test administrators are used (the HSPT uses neither), the cost is about 100 times the cost of multiple choice testing. Given such estimates, the constructed response items must give additional, meaningful information if their cost is to be justified.

Directions for Further Research

Further research into the constructs measured by each format, and their possible effects on teaching and learning, would be useful to educators weighing the costs and benefits of the different formats. As noted, the constructs measured by the constructed response and multiple choice items are highly correlated. Though the constructed response items do add information to the scores, additional multiple choice items might

add more information at a cheaper cost. If the constructed response items are intended to reinforce certain instructional methods, it would be reasonable to expect research, perhaps a survey, to explore this relationship. If the constructed response section is intended to boost teacher/student/community perceptions and attitudes toward the test, it would make sense to study perceptions and attitudes. Such findings could help in determining the direction of future test development.

A large part of the ethnic difference in performance in science was due to between-school differences. More school-level factors could be entered into the model to examine some factors which might explain these school differences. School climate, curriculum, and community factors are likely associated with both ethnicity and school performance. After measures of these factors were included, the school's ethnic composition would likely account for little additional variation. School climate and curriculum are also potentially changeable, so knowing more about how these factors relate to student achievement would be useful to policy-makers.

Further research using other tests and student populations will help to reveal to which situations the results of this study generalize.

APPENDIX

RESULTS FROM THE SECOND YEAR OF THE SCIENCE TEST

APPENDIX

RESULTS FROM THE SECOND FORM OF THE SCIENCE TEST

Table A1 - Participants

	White Females	Black Females	White Males	Black Males
High Stakes	1454	241	1311	211
Low Stakes	519	68	547	57

Average scale score in tested population (Blacks and Whites only): 387.06 (sd = 42.67)

Average scale score in selected schools (Blacks and Whites only): 389.58 (sd = 42.01)

Table A2 - Student-Level Predictors for Log-Odds of Response, Science Year 2

	Coefficient	Standard Error	Approx. T-ratio	d.f	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	2.800651	0.095298	29.388	37	0.000
For STAKES (S), β_{01k}					
INTERCEPT3, γ_{010}	0.933603	0.074453	12.540	4369	0.000
For ETHNICITY (E), β_{02k}					
INTERCEPT3, γ_{020}	-0.810397	0.197113	-4.111	4369	0.000
For GENDER (G), β_{03k}					
INTERCEPT3, γ_{030}	-0.533162	0.069176	-7.707	4369	0.000
For S X E, β_{04k}					
INTERCEPT3, γ_{040}	0.794302	0.221220	3.591	4369	0.001
For S X G, β_{05k}					
INTERCEPT3, γ_{050}	-0.010617	0.138337	-0.077	4369	0.939
For E X G, β_{06k}					
INTERCEPT3, γ_{060}	-0.295909	0.198696	-1.489	4369	0.136
For S X G X E, β_{07k}					
INTERCEPT3, γ_{070}	0.001481	0.396929	0.004	4369	0.997

Table A3 - Log-Odds of Response on Science CR Items, Full Model Year 2

	Coefficient	Standard Error	Approx. T-ratio	d.f	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	2.891245	0.072705	39.767	36	0.000
MINORITY, γ_{001}	-1.066216	0.307136	-3.471	36	0.002
For STAKES (S) , β_{01k}					
INTERCEPT3, γ_{010}	0.716872	0.100991	7.098	4368	0.000
MINORITY, γ_{011}	0.988755	0.683999	1.446	4368	0.148
For ETHNICITY (E), β_{02k}					
INTERCEPT3, γ_{020}	-0.267060	0.209546	-1.274	4368	0.203
MINORITY, γ_{021}	-0.628945	0.407656	-1.543	4368	0.123
For GENDER (G), β_{03k}					
INTERCEPT3, γ_{030}	-0.608106	0.097247	-6.253	4368	0.000
MINORITY, γ_{031}	-1.090945	0.537663	-2.029	4368	0.042
For S x E, β_{04k}					
INTERCEPT3, γ_{040}	-0.999135	0.363708	-2.747	4368	0.006
MINORITY, γ_{041}	1.521008	0.917019	1.659	4368	0.097
For S X G, β_{05k}					
INTERCEPT3, γ_{050}	0.335921	0.161174	2.084	4368	0.037
MINORITY, γ_{051}	-0.210093	1.443877	-0.146	4368	0.885
For E X G, β_{06k}					
INTERCEPT3, γ_{060}	-0.410597	0.323266	-1.270	4368	0.204
MINORITY, γ_{061}	1.591441	0.851798	1.868	4368	0.061
For S.X GX E, β_{07k}					
INTERCEPT3, γ_{070}	2.444570	0.575440	4.248	4368	0.000
MINORITY, γ_{071}	-3.264864	1.735777	-1.881	4368	0.060

Table A4 - Log-Odds of Response on Science CR Items, Final Model Year 2

	Coefficient	Standard Error	Approx. T-ratio	d.f	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	2.844423	0.065017	43.749	36	0.000
MINORITY, γ_{001}	-1.264149	0.264041	-4.788	36	0.000
For STAKES, β_{01k}					
INTERCEPT3, γ_{010}	0.856606	0.092626	9.248	4368	0.000
MINORITY, γ_{011}	1.052756	0.155070	6.789	4368	0.000
For ETHNICITY, β_{02k}					
INTERCEPT3, γ_{020}	-0.489279	0.144558	-3.385	4369	0.001
For GENDER, β_{03k}					
INTERCEPT3, γ_{030}	-0.500459	0.078295	-6.392	4368	0.000
MINORITY, γ_{031}	-0.374395	0.124985	-2.996	4368	0.003

Table A5 - Predicted Log-Odds, Probability (Rate), and Odds of Response, Year 2

	log-odds	probability	probability ratio	odds	odds ratio
population average					
gender			1.03		1.65
girls	3.08	0.96		21.78	
boys	2.58	0.93		13.21	
stakes			1.05		2.36
high stakes	3.27	0.96		26.38	
low stakes	2.42	0.92		11.20	
ethnicity			0.97		0.61
Black	2.42	0.92		11.29	
White	2.91	0.95		18.42	

Table A6- Schools with High and Low Minority Enrollment, Science Year 2

5% Minority	log-odds	probability	probability ratio	odds	odds ratio
population average	3.12	0.96		22.65	
gender			1.02		1.55
girls	3.33	0.97		27.85	
boys	2.89	0.95		18.00	
stakes			1.03		1.97
high stakes	3.46	0.97		31.79	
low stakes	2.78	0.94		16.14	
ethnicity			0.98		0.61
Black	2.70	0.94		14.88	
White	3.19	0.96		24.27	

90% Minority	log-odds	probability	probability ratio	odds	odds ratio
population average	1.74	0.85		5.70	
gender			1.12		2.13
girls	2.10	0.89		8.15	
boys	1.34	0.79		3.83	
stakes			1.28		4.82
high stakes	2.53	0.93		12.52	
low stakes	0.95	0.72		2.60	
ethnicity			0.92		0.61
Black	1.32	0.79		3.75	
White	1.81	0.86		6.11	

Table A7 - Student-Level Predictors for Ability Scores, Science Year 2

	Coefficient	Standard Error	Approx. T-ratio	d.f	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	-0.368311	0.032772	-11.239	37	0.000
For STAKES (S), β_{01k}					
INTERCEPT3, γ_{010}	0.628454	0.037958	16.557	4331	0.000
For ETHNICITY (E), β_{02k}					
INTERCEPT3, γ_{020}	-0.792693	0.104973	-7.551	4331	0.000
For GENDER (G), β_{03k}					
INTERCEPT3, γ_{030}	0.090958	0.028244	3.220	4331	0.002
For S X E, β_{04k}					
INTERCEPT3, γ_{040}	0.015578	0.128680	0.121	4331	0.904
For S X G, β_{05k}					
INTERCEPT3, γ_{050}	0.074922	0.054425	1.377	4331	0.169
For E X G, β_{06k}					
INTERCEPT3, γ_{060}	-0.148572	0.064037	-2.320	4331	0.020
For S X G X E, β_{07k}					
INTERCEPT3, γ_{070}	0.025874	0.129166	0.200	4331	0.841
For FORMAT slope, π_{1jk}					
For INTERCEPT2, β_{10k}					
INTERCEPT3, γ_{100}	-0.303127	0.015753	-19.242	37	0.000
For STAKES, β_{11k}					
INTERCEPT3, γ_{110}	0.590857	0.024124	24.493	4407	0.000
For ETHNICITY, β_{12k}					
INTERCEPT3, γ_{120}	0.114266	0.037900	3.015	4407	0.003
For GENDER, β_{13k}					
INTERCEPT3, γ_{130}	-0.248776	0.023390	-10.636	4407	0.000
For S X E, β_{14k}					
INTERCEPT3, γ_{140}	-0.040230	0.071488	-0.563	4407	0.573
For S X G, β_{15k}					
INTERCEPT3, γ_{150}	-0.047652	0.041683	-1.143	4407	0.253
For E X G, β_{16k}					
INTERCEPT3, γ_{160}	0.123967	0.076168	1.628	4407	0.103
For S X G X E, β_{17k}					
INTERCEPT3, γ_{170}	-0.050973	0.094102	-0.542	4407	0.588

Table A8 - Predicted Scores, Controlling School Minority Composition, Year 2

	Coefficient	Standard Error	Approx. T-ratio	d.f	P-value
For INTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	-0.370712	0.040003	-9.267	36	0.000
MINORITY, γ_{001}	-0.295758	0.206037	-1.435	36	0.160
For STAKES, β_{01k}					
INTERCEPT3, γ_{010}	0.630912	0.037307	16.911	4330	0.000
MINORITY, γ_{011}	-0.021979	0.176277	-0.125	4330	0.901
For ETHNICITY, β_{02k}					
INTERCEPT3, γ_{020}	-0.755193	0.170286	-4.435	4330	0.000
MINORITY, γ_{021}	0.110763	0.266260	0.416	4330	0.677
For GENDER, β_{03k}					
INTERCEPT3, γ_{030}	0.124936	0.035782	3.492	4330	0.001
MINORITY, γ_{031}	0.042350	0.267658	0.158	4330	0.875
For E x G, β_{04k}					
INTERCEPT3, γ_{040}	-0.044665	0.130526	-0.342	4330	0.732
MINORITY, γ_{041}	-0.204481	0.403585	-0.507	4330	0.612
For FORMAT slope, π_{1jk}					
For INTERCEPT2, β_{10k}					
INTERCEPT3, γ_{100}	-0.301751	0.015070	-20.023	36	0.000
MINORITY, γ_{101}	-0.159088	0.090054	-1.767	36	0.085
For STAKES, β_{11k}					
INTERCEPT3, γ_{110}	0.592530	0.024506	24.178	4406	0.000
MINORITY, γ_{111}	0.010234	0.060442	0.169	4406	0.866
For ETHNICITY, β_{12k}					
INTERCEPT3, γ_{120}	0.156582	0.067906	2.306	4406	0.021
MINORITY, γ_{121}	0.062494	0.134940	0.463	4406	0.643
For GENDER, β_{13k}					
INTERCEPT3, γ_{130}	-0.267795	0.020918	-12.802	4406	0.000
MINORITY, γ_{131}	0.120736	0.053582	2.253	4406	0.024

Table A9 - The Final Model for Science Ability Scores, Year 2

	Coefficient	Standard Error	Approx. T-ratio	d.f	P-value
ForINTERCEPT1, π_{0jk}					
For INTERCEPT2, β_{00k}					
INTERCEPT3, γ_{000}	-0.368950	0.032495	-11.354	37	0.000
For STAKES, β_{01k}					
INTERCEPT3, γ_{010}	0.629250	0.037422	16.815	4331	0.000
For ETHNICITY, β_{02k}					
INTERCEPT3, γ_{020}	-0.796182	0.103219	-7.714	4331	0.000
For GENDER, β_{03k}					
INTERCEPT3, γ_{030}	0.118441	0.028818	4.110	4330	0.000
MINORITY, γ_{031}	-0.158953	0.048068	-3.307	4330	0.001
For FORMAT slope, π_{1jk}					
For INTERCEPT2, β_{10k}					
INTERCEPT3, γ_{100}	-0.298535	0.015854	-18.830	36	0.000
MINORITY, γ_{101}	-0.107970	0.069178	-1.561	36	0.127
For STAKES, β_{11k}					
INTERCEPT3, γ_{110}	0.592874	0.023961	24.743	4407	0.000
For ETHNICITY, β_{12k}					
INTERCEPT3, γ_{120}	0.163372	0.055858	2.925	4407	0.004
For GENDER, β_{13k}					
INTERCEPT3, γ_{130}	-0.267749	0.020878	-12.825	4406	0.000
MINORITY, γ_{131}	0.120501	0.053250	2.263	4406	0.024

Table A10 - Average Predicted Scores in Science, Year 2

	multiple choice		constructed response		mean	
	mean	sd	mean	sd	mean	sd
average	-0.22	0.85	-0.52	0.89	-0.37	0.80
high stakes	-0.05	0.82	-0.06	0.84	-0.05	0.77
low stakes	-0.39	0.91	-0.98	0.76	-0.68	0.77
Blacks	-0.97	0.70	-1.13	0.74	-1.05	0.64
Whites	-0.10	0.81	-0.42	0.87	-0.26	0.76
Boys	-0.09	0.91	-0.53	0.92	-0.31	0.84
Girls	-0.34	0.79	-0.51	0.86	-0.42	0.75

	5% minority			90% minority		
	MC	CR	mean	MC	CR	mean
average	-0.23	-0.51	-0.37	-0.18	-0.55	-0.37
high stakes	-0.06	-0.05	-0.05	-0.02	-0.09	-0.05
low stakes	-0.40	-0.97	-0.68	-0.35	-1.02	-0.68
Blacks	-0.98	-1.12	-1.05	-0.94	-1.17	-1.05
Whites	-0.11	-0.41	-0.26	-0.06	-0.45	-0.26
Boys	-0.08	-0.51	-0.29	-0.13	-0.60	-0.36
Girls	-0.37	-0.51	-0.44	-0.23	-0.52	-0.37

The standard deviation of ability scores in the total sample was 0.87; this standard deviation was used to calculate the standardized differences in Table A11.

Table A11 - Standardized Within-School Effects in Science, Year 2

	standardized difference		
	multiple choice	constructed response	average
stakes	0.38	1.06	0.72
ethnicity	-1.01	-0.82	-0.91
gender	0.29	-0.02	0.14

The standardized ethnic group differences across the total sample were -1.03 for multiple choice, -.85 for constructed response, and -.95 for the average.

LIST OF REFERENCES

LIST OF REFERENCES

Abdel-fattah, A. (1994, April). Comparing BILOG and LOGIST estimates for normal, truncated normal, and beta ability distributions. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 374158)

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. Personnel Psychology, 43, 695-716.

Asher, S. R. (1979). Influence of topic interest on black children's and white children's reading comprehension. Child Development, 50, 686-690.

Asher, S. R. (1980). Topic interest and children's reading comprehension. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), Theoretical issues in reading comprehension (pp. 525-534). Hillsdale, NJ: Lawrence Erlbaum Associates.

Badger, E. (1989). On their own: Student response to open-ended tests in math. Quincy, MA: Massachusetts State Department of Education. (ERIC Document Reproduction Service No. ED 317573)

Badger, E. (1995). The effects of expectations on achieving equity in state-wide testing: Lessons from Massachusetts. In M. T. Nettles & A. L. Nettles (Eds.), Equity and excellence in educational testing and assessment (pp. 289-308). Boston: Kluwer Academic.

Badger, E., & Thomas, B. (1989a). On their own: Student response to open-ended tests in reading. Quincy, MA: Massachusetts State Department of Education. (ERIC Document Reproduction Service No. ED 317573)

Badger, E., & Thomas, B. (1989b). On their own: Student response to open-ended tests in science. Quincy, MA: Massachusetts State Department of Education. (ERIC Document Reproduction Service No. ED 317573)

Badger, E., Thomas, B., & McCormack, E. (1990). Beyond paper and pencil. Boston: Massachusetts Educational Assessment Program, Massachusetts State Department of Education. (ERIC Document Reproduction Service No. ED 336400)

Bennett, R.E., Rock, D.A., & Wang, M. (1991). Equivalence of free-response and multiple choice items. Journal of Educational Measurement, 28, 77-92.

Bernstein, M. R. (1955). Relationship between interest and reading comprehension. Journal of Educational Research, 49, 283-288.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement, 27, 165-174.

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. Educational Measurement: Issues and Practice, 14 (4), 21-24.

Bridgeman, B. (1989). Comparative validity of multiple choice and free-response items on the advanced placement examination in biology (College Board Report No. 89-2). (ERIC Document Reproduction Service No. ED 308 228)

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. Journal of Educational Measurement, 31, 37-50.

Brown, S. M., & Walberg, H. J. (1993). Motivational effects of test scores of elementary students. Journal of Educational Research, 86, 133-136.

Burke, P. (1991). You can lead adolescents to a test but you can't make them try. Final report (Contract No. OTA-H3-6110.0). Washington, D.C.: Office of Technology Assessment. (ERIC Document Reproduction Service No. ED 378221)

Bryk, A., & Raudenbush, S. (1992). Hierarchical linear models. Newbury Park, CA: Sage Publications.

Bryk, A., Raudenbush, S., Congdon, R. (1996). HLM4: Hierarchical linear & nonlinear modeling [Computer software]. Chicago: Scientific Software International.

Chandler, T. A., & Spies, C. J. (1981). Attribution as predictor of expectancy in three component exams. Teaching of Psychology, 8, 174-175.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. Journal of Applied Psychology, 82, 143-159.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. Review of Educational Research, 58, 438-481.

Curran, M. T., & Harich, K. R. (1993). Performance attributions: Effects of mood and involvement. Journal of Educational Psychology, 85, 605-609.

Darling-Hammond, L. (1995). Equity issues in performance-based assessment. In M. T. Nettles & A. L. Nettles (Eds.), Equity and excellence in educational testing and assessment (pp. 89-114). Boston: Kluwer Academic.

DeMars, C. (in press). Gender differences in mathematics and science on a high school proficiency exam. Applied Measurement in Education.

Drasgow, F., Levine, M.V., Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. British Journal of Mathematical and Statistical Psychology, 38, 67-86.

Dreeban, R., & Gamoran, A. (1986). Race, instruction, and learning. American Sociological Review, 51, 660-669.

Feinberg, L. (1990). Multiple-choice and its critics. The College Board Review, (157), 12-17, 30-31.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193-202.

Freund, D. S., & Rock, D. A. (1992, April). A preliminary investigation of pattern-marking in 1990 NAEP data. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 347189)

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. Applied Measurement in Education, 7, 323-342.

Gerber, S. (1996, April). Self-efficacy, item difficulty, and persistence in constructed-response tests. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. Review of Educational Research, 58, 47-77.

Hill, K. T. (1980). Eliminating motivational causes of test bias. Final report, October 1, 1976 through March 31, 1980. Urbana, IL: Illinois University, Institute for Child Behavior and Development. (ERIC Document Reproduction Service No. ED 196936)

Jennings, E. E. (1953). The motivation factor in testing supervisors. Journal of Applied Psychology, 37, 168-169.

Johnson, S. T. (1995). Visions of equity in national assessment. In M. T. Nettles, & A. L. Nettles (Eds.), Equity and excellence in educational testing and assessment (pp. 343-366). Boston: Kluwer Academic.

Karmos, A. H., & Karmos, J. S. (1984). Attitudes toward standardized achievement tests and their relation to achievement test performance. Motivation and Evaluation in Counseling and Development, 17, 56-66.

Kiplinger, V. L., & Linn, R. L. (1992, April). Raising the stakes of test administration: The impact on student performance on NAEP. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles. (ERIC Document Reproduction Service No. ED 378221)

Klein, S.P., Jovanovic, J., Stecher, B.M., McCaffrey, D., Shavelson, R.J., Haertel, E., Solano-Flores, G., and Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. Educational Evaluation and Policy Analysis, 19, 83-97.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20 (8), 15-21.

Lord, F. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. Journal of Educational Measurement, 23, 157-162.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23 (2), 13-23.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57-75.

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. Journal of Counseling Psychology, 38, 30-38.

Muraki, E., & Bock, R.D. (1997). PARSCALE 3.3 [Computer software]. Chicago: Scientific Software International.

Murphy, R.J.L., 1982. Sex differences in GCE examination entry statistics and success rates. Educational Studies, 6, 169-178.

Olsen, S. A., & Wilson, K. (1991). A follow-up of suspect sophomores' scores on the COMP test. Kirksville, MO: Northeast Missouri State University. (ERIC Document Reproduction Service No. ED 339735)

Paris, S. G., Lawton, T. A., & Turner, J. C. (1992). Reforming achievement testing to promote students' learning. In C. Collins & J. M. Mangieri (Eds.), Teaching thinking: An agenda for the 21st century (pp. 223-241). Hillsdale, NJ: Lawrence Erlbaum Associates.

Paris, S. G., Turner, J. C., Lawton, T. A., & Roth, J. L. (1991). A developmental perspective on standardized achievement testing. Educational Researcher, 20, 12-20.

Plass, J. A., & Hill, K. T. (1986). Children's achievement strategies and test performance: The role of time pressure, evaluation anxiety, and sex. Developmental Psychology, 22, 31-36.

Pomplun, M. (1997). State assessment and instructional change: A path model analysis. Applied Measurement in Education, 10, 217-234.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. Journal of Educational Measurement, 27, 133-144.

Roeber, E. D. (1995). Using new forms of assessment to assist in achieving student equity: Experiences of the CCSSO State Collaborative on Assessment and Student Standards. In M. T. Nettles & A. L. Nettles (Eds.), Equity and excellence in educational testing and assessment (pp. 265-288). Boston: Kluwer Academic.

Rothe, H. F. (1947). Distribution of test scores in industrial employees and applicants. Journal of Applied Psychology, 31, 480-483.

Schmidt, F. L., Greenthal, A. L., Berner, J. G., Hunter, J. E., & Seaton, F. W. (1977). Job sample vs. paper-and-pencil trades and technical tests: Adverse impact and examinee attitudes. Personnel Psychology, 30, 187-197.

Schmitt, A. P., Mazzeo, J., & Bleistein, C. (1991, April). Are gender differences between Advanced Placement multiple choice and constructed response sections a

function of multiple choice DIF? Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago.

Segall, D. O. (1997, March). The effects of motivation on equating adaptive and conventional tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 21 (4), 22-27.

Shepard, L. (1993). Evaluating test validity. Review of Research in Education, 19, 405-450.

Smith, R. E., & Smoll, F. L. (1990). Sport performance anxiety. In H. Leitenberg (Ed.), Handbook of social and evaluation anxiety (pp. 417-454). New York: Plenum Press.

Stanley, G., King, M., & Glass, J. (1989). Stress arousal and performance. In J. P. Forgas & J. M. Innes (Eds.), Recent advances in social psychology: An international perspective (pp. 229-234). Amsterdam: Elsevier Science Publishers.

Stecher, B. (1995, April). The cost of performance assessment in science: The RAND perspective. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Stecher, B., & Hamilton, E. G. (1994, April). Portfolio assessment in Vermont, 1992-93: The teachers' perspective on implementation and impact. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans

Stevens, K. (1979). The effect of topic interest on the reading comprehension of higher ability students. Journal of Educational Research, 73, 365-368.

Stocking, M. L. (1989). Empirical estimation errors in item response theory as a function of test properties. NJ: ETS. (ERIC Document Reproduction Service No. ED 395027)

Sundre, D. L. (1996). The role of examinee motivation in assessment: Celebrity, scene stealer, or cameo? Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Taylor, C., & White, K. R. (1981, April). Effects of reinforcement and training on Title I students' group standardized test performance. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles. (ERIC Document Reproduction Service No. ED 206655)

Thomas, B. (1989). On their own: Student response to open-ended tests in social studies. Quincy, MA: Massachusetts State Department of Education. (ERIC Document Reproduction Service No. ED 317573)

Tobias, S. (1985). Test anxiety: Interference, defective skills, and cognitive capacity. Educational Psychologist, 20, 135-142.

Uguroglu, M. E., & Walberg, H. J. (1979). Motivational achievement: A quantitative synthesis. American Educational Research Journal, 16, 375-390.

Wainer, H. (1993). Measurement problems. Journal of Educational Measurement, 30, 1-21.

Wainer, H., & Lukhele, R. (1997). Managing the influence of DIF from big items: The 1988 Advanced Placement History Test as an example. Applied Measurement in Education, 10, 201-215.

Winfield, L. F. (1995). Performance-based assessments: Contributor or detractor to equity? In M. T. Nettles & A. L. Nettles (Ed.), Equity and excellence in educational testing and assessment (pp. 221-241). Boston: Kluwer Academic.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. Applied Measurement in Education, 8, 227-242.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. Applied Measurement in Education, 8, 341-351.

Wolf, L. F., Smith, J. K., & DiPaulo, T. (1996, April). The effects of test specific motivation and anxiety on test performance. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187-213.

Yen, W.M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.